Trabajo Fin de Máster

Análisis de datos funcionales para clasificación con datos de espectroscopía

Mª Dolores de la Fuente Blanco

Máster en Técnicas Estadísticas

Curso 2023-2024

Propuesta de Trabajo Fin de Máster

Título en galego: Análise de datos funcionais para clasificación con datos de espectroscopía
Título en español: Análisis de datos funcionales para clasificación con datos de espectroscopía
English title: Functional data analysis for classification with spectroscopy data
Modalidad: Modalidad A
Autor/a: Mª Dolores de la Fuente Blanco, Universidad de Santiago de Compostela
Director/a: Manuel Febrero Bande, Universidad de Santiago de Compostela; Beatriz Pateiro López, Universidad de Santiago de Compostela
Breve resumen del trabajo: Este trabajo trata la detección de cáncer de mama y colon como un problema de clasificación binario. Se cuenta con una base de datos de espectros infrarrojos de suero sanguíneo, lo que aporta información sobre la composición molecular de la muestra. Debido a la naturaleza intrínsecamente funcional de los mismos, se aplicarán técnicas de análisis de datos funcionales.

Don/doña Manuel Febrero Bande, Profesor Catedrático de la Universidad de Santiago de Compostela, don/doña Beatriz Pateiro López, Profesora Titular de la Universidad de Santiago de Compostela, informan que el Trabajo Fin de Máster titulado

Análisis de datos funcionales para clasificación con datos de espectroscopía

fue realizado bajo su dirección por don/doña Mª Dolores de la Fuente Blanco para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Santiago de Compostela, a 3 de junio de 2024.

El/la director/a:
Don/doña Manuel Febrero Bande

El/la director/a:
Don/doña Beatriz Pateiro López

El/la tutor/a:
Don/doña

El/la tutor/a:
Don/doña

El/la autor/a:
Don/doña Mª Dolores de la Fuente Blanco

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el/la autor/a declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.

- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,... sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Agradecimientos

A mis tutores, Manuel y Beatriz, por acompañarme en este trabajo, por su tiempo y su implicación; sin los que este trabajo no hubiera sido posible.

A mis compañeros de Activa Biotech, por dos años de gran aprendizaje. En especial a Samu, por todas las horas de laboratorio para conseguir esta base de datos y a José Ángel, por darme la oportunidad de participar en este apasionante proyecto y ser ejemplo de trabajo incansable.

Al Biobanco del Principado de Asturias (PT20/0161), cofinanciado por el Servicio de Salud del Principado de Asturias, el Instituto de Salud Carlos III y la Fundación Bancaria Cajastur e integrado en la Red Nacional de Biobancos y Biomodelos y al Centro Comunitario de Sangre y Tejidos de Asturias por su colaboración.

A mi familia, por darme el apoyo, la educación y los valores de esfuerzo, disciplina y persistencia necesarios para llegar hasta aquí. En especial a mi hermana, Alba, por ser mi mejor modelo a seguir.

A mis amigos, a los que admiro enormemente y me inspiran y motivan cada día día, por el apoyo incondicional acompañándome en este camino, la confianza y los buenos momentos que hemos compartido y siempre recordaré.

Índice general

Resumen xi

Prefacio xiii

1. Motivación y presentación de la base de datos 1
- 1.1. Motivación del problema . . . 1
- 1.2. Grupo de cáncer de mama . . . 3
- 1.3. Grupo de cáncer de colon . . . 3
- 1.4. Grupo de control . . . 4
- 1.5. Espectroscopía infrarroja . . . 4

2. Introducción a datos funcionales 7
- 2.1. Motivación . . . 7
- 2.2. Primeras definiciones . . . 8
  - 2.2.1. Representación en una base de Componentes Principales Funcional . . . 9
- 2.3. Medidas de localización para datos funcionales . . . 9
- 2.4. Análisis de las Componentes Principales Funcionales . . . 11
- 2.5. ANOVA Funcional . . . 12
- 2.6. Clasificadores para Datos Funcionales . . . 13
  - 2.6.1. Métricas de evaluación . . . 14

3. Análisis exploratorio de los datos 17
- 3.1. Medidas de localización y dispersión . . . 17
- 3.2. Detección de atípicos . . . 19
- 3.3. FPCA . . . 22

4. Modelos de clasificación 25
- 4.1. ANOVA . . . 25
  - 4.1.1. Factor grupo . . . 25
  - 4.1.2. Factor edad . . . 28
  - 4.1.3. Factor sexo . . . 28
  - 4.1.4. ANOVA multifactorial . . . 31
- 4.2. Modelos de clasificación: cáncer de mama vs referencia . . . 33
- 4.3. Modelos de clasificación: cáncer de colon vs referencia . . . 38

Conclusiones **45**

A. Código del Análisis Exploratorio **47**

B. Código ANOVA y clasificación **49**
B.1. ANOVA 49
B.2. Clasificación 51
B.3. Funciones auxiliares 55

Bibliografía **58**

Resumen

Resumen en español

Este trabajo trata la detección de cáncer de mama y colon como un problema de clasificación binario. Se cuenta con una base de datos de espectros infrarrojos de suero sanguíneo, lo que aporta información sobre la composición molecular de la muestra. Debido a la naturaleza intrínsecamente funcional de los mismos, se aplicarán técnicas de análisis de datos funcionales.

Una vez motivado el problema, se presentan las nociones y herramientas de análisis de datos funcionales necesarias para el estudio. A continuación, se realiza un análisis exploratorio de la base de datos, representando medidas de localización central, estudiando la existencia de atípicos y realizando un análisis de componentes principales funcional.

Por último, se ajustan distintos tipos de modelos de clasificación, siendo kNN el que mejores resultados obtuvo en distintas métricas de evaluación para el caso de cáncer de mama contra el grupo de referencia y el análisis del discriminante cuadrático, QDA, en el caso de grupo de pacientes de cáncer de colon contra individuos del grupo de referencia.

Finalmente, se obtuvieron resultados prometedores que establecen una base sólida para el uso de la espectroscopía infrarroja en la detección del cáncer. Estos subrayan la importancia de ampliar la base de datos y explorar nuevas técnicas analíticas para mejorar la escalabilidad y generalización de los resultados, lo que es esencial para futuras investigaciones y aplicaciones clínicas.

English abstract

This work addresses breast and colon cancer detection as a binary classification problem. A database of infrared spectra of blood serum is provided, which gives information on the molecular composition of the sample. Due to their intrinsically functional nature, functional data analysis techniques will be applied.

Once the problem has been motivated, the notions and tools of functional data analysis that are required for the study are presented. Then, an exploratory analysis of the database is performed, representing measures of central location, studying the existence of outliers and performing a functional principal component analysis.

Finally, different types of classification models are fitted, being kNN the best performer in different evaluation metrics for the case of breast cancer against the reference group and quadratic discriminant analysis, QDA, in the case of colon cancer patients group against individuals from the reference group.

Finally, promising results were obtained that establish a solid basis for the use of infrared spectroscopy in cancer detection. These underline the importance of expanding the database and exploring new analytical techniques to improve the scalability and generalisability of the results, which is essential for future research and clinical applications.

Prefacio

La Estadística es una rama de las Matemáticas que permite recolectar información para luego analizarla y extraer de ella conclusiones relevantes. Con el auge de los datos en la era de la información han aparecido nuevos tipos de datos, como intervalos, funciones, conjuntos difusos, distribuciones, etc, lo que ha supuesto una auténtica revolución para la estadística. En particular, los datos funcionales pueden encontrarse comúnmente en aplicaciones diversas en numerosos campos de estudio, como Química, Bioestadística, Ingeniería, etc. Estos datos pueden proceder, por ejemplo, de la monitorización de niveles de presión sanguínea o glucosa para pacientes, perfiles diarios de temperatura, sustancias contaminantes u otras variables meteorológicas, observaciones de cualidades propiamente funcionales.

Los métodos estadísticos tradicionales fallan al trabajar con bases de datos funcionales. De hecho, al trabajar con una muestra de curvas discretizadas, dos problemas cruciales aparecen. El primero es la baja proporción entre el tamaño muestral y el número de variables (donde cada variable real se corresponde con un punto discretizado). El segundo es la existencia de fuertes correlaciones entre variables, convirtiendo el problema en uno mal condicionado en el contexto del análisis multivariante, con problemas de inestabilidad numérica al ajustar modelos resultando en poca fiabilidad e interpretación de los resultados obtenidos del mismo. La adaptación de la metodología para tratar este tipo de datos y el desarrollo de técnicas *ad hoc* no son triviales debido a las peculiaridades de los distintos espacios funcionales que nos podemos encontrar. Esto lleva a una necesidad de desarrollar métodos y modelos estadísticos que tengan en cuenta la estructura funcional de este tipo de datos. Este trabajo se centrará en el estudio del problema de clasificación para datos funcionales.

Una vez motivado el estudio de los datos funcionales, en el **Capítulo 1** se presenta la base de datos sobre la que se centrará el análisis. Esta fue obtenida a partir de muestras de suero sanguíneo de pacientes con cáncer del biobanco del Principado de Asturias, así como de individuos sanos de referencia.

En el **Capítulo 2** se revisan las nociones teóricas básicas de datos funcionales. Se exponen las singularidades de los espacios funcionales y se presentan la noción de variable aleatoria funcional, que formaliza matemáticamente el proceso de generación aleatoria de datos funcionales. También se presentarán formalmente las herramientas estadísticas utilizadas en los siguientes capítulos.

En el **Capítulo 3** se realiza un análisis exploratorio de los datos. Se calculan medidas de localización y dispersión y se analiza la existencia de posibles datos atípicos. Además, se realiza un análisis de componentes principales funcional que ayuda a identificar patrones, evaluar la importancia de las variables y explorar la estructura subyacente de los datos. El código de programación puede consultarse en los anexos.

En el **Capítulo 4** se estudia la influencia de los factores grupo, sexo y edad en el problema a

través de ANOVA unifactorial y multifactorial. A continuación, se presentan los resultados de evaluar distintos clasificadores binarios: multivariantes, basados en estimación kernel, en regresión binaria y en profundidades.

Finalmente, en las **Conclusiones** se resumen los aspectos y los resultados más importantes del trabajo, así como sus principales limitaciones.

Capítulo 1

Motivación y presentación de la base de datos

En este capítulo se motivará el problema de clasificación a tratar en este trabajo. También se presentará la base de datos de espectros infrarrojos con la que se cuenta, exponiendo sus grupos y características principales.

1.1. Motivación del problema

El cáncer es un problema de salud crítico en todo el mundo, ya que es una de las principales causas de morbilidad y mortalidad. La amplia variedad de cánceres destaca la necesidad de aumentar los esfuerzos globales para combatir la enfermedad. Especialmente, el cáncer de mama se presenta como el más común en las mujeres, así como la principal causa de muerte relacionada con el cáncer entre ellas, representando el 24.5 % de todos los casos de cáncer y 15.5 % de las muertes por cáncer, casi 700000 muertes anuales. Considerando ambos sexos, el cáncer de colon ocupa el cuarto lugar en incidencia global y el quinto en mortalidad, representando aproximadamente 550000 muertes anuales en el mundo. De todos los pacientes con cáncer de colon, el 25 % son diagnosticados en una etapa avanzada, lo que se asocia con un mal pronóstico y una tasa de supervivencia global a cinco años de solo 14 % [Sung et al., 2021]. La detección temprana es crucial en el manejo del cáncer, ya que mejora significativamente las posibilidades de tratamiento exitoso y supervivencia [Jayasinghe et al., 2023].

Existe una gran necesidad de nuevos métodos de detección que puedan involucrar efectivamente a la población objetivo, lo que conduciría a una detección temprana y al inicio del tratamiento antes de que se manifiesten los síntomas clínicos. Este enfoque no solo reduciría las tasas de mortalidad, sino que también mejoraría la calidad de vida de las personas afectadas al proporcionar un pronóstico más favorable. Además, la intervención temprana permite opciones de tratamiento menos agresivas, especialmente cuando la enfermedad se detecta en su fase preclínica.

Los métodos de detección de cáncer actuales a menudo no cumplen con las demandas tanto de los servicios de salud como de los consumidores, lo que lleva a tasas de participación bajas en los programas de detección y a una detección tardía. La detección temprana es necesaria para reducir tanto la incidencia como la mortalidad. Sin embargo, hay diferencias significativas entre los estudios

en las estrategias de implementación y detección[Navarro et al., 2017].

El principal inconveniente de los métodos de detección son los resultados falsos positivos o falsos negativos. Los falsos positivos pueden generar costos adicionales y angustia emocional para los participantes, lo que potencialmente afecta su futura participación en los programas de detección. Por otro lado, los resultados falsos negativos pueden retrasar la detección de lesiones avanzadas, proporcionando una falsa tranquilidad a los participantes. Para el cáncer de mama, la mamografía se considera el método de detección estándar. Sin embargo, además de la incomodidad que muchas mujeres experimentan debido a la compresión mamaria necesaria para una mamografía óptima, la mamografía presenta varios riesgos. Estos incluyen resultados falsos positivos y falsos negativos, sobrediagnóstico que conduce a un tratamiento excesivo y el potencial de cánceres inducidos por radiación [Warner, 2011]. Para la detección del cáncer de colon, están disponibles varias modalidades, principalmente pruebas basadas en heces (por ejemplo, prueba de sangre oculta en heces, FIT) y exámenes endoscópicos visuales (por ejemplo, colonoscopia). La principal limitación asociada con la colonoscopia es su naturaleza invasiva, lo que puede reducir las tasas de participación entre la población. Además, estudios recientes han subrayado preocupaciones sobre resultados falsos positivos y falsos negativos en las pruebas inmunocromatográficas fecales (FIT) empleadas para la detección del cáncer de colon, lo que requiere una colonoscopia de seguimiento en caso de un resultado positivo [Wong et al., 2015].

En este trabajo se estudia la capacidad de la espectroscopía infrarroja combinada con técnicas de aprendizaje estadístico como método de detección temprana de cáncer de mama y colon. Esta tecnología usa energía del infrarrojo medio modulada para analizar una muestra, ayudando a comprender la estructura de las moléculas individuales y la composición de las mezclas moleculares. Además, ha demostrado ser una forma mínimamente invasiva, rápida, efectiva y económica de detectar de manera temprana varias enfermedades [Huber et al., 2021, Guang et al., 2020, Paraskevaidi et al., 2017]. Estas características la convierten en una propuesta sólida para realizar cribados eficientes.

Las muestras y datos de donantes de pacientes de cáncer incluidos en este estudio fueron proporcionados por el Biobanco del Principado de Asturias (PT20/0161), integrado en la Red Nacional de Biobancos y Biomodelos, con la aprobación de sendos Comités Ético y Científico, mientras que aquellas del grupo de referencia fueron proporcionadas por el Centro Comunitario de Sangre y Tejidos de Asturias. Todas las muestras fueron medidas en el laboratorio de Espectroscopía Molecular y XPS en los Servicios Científico-Técnicos "Severo Ochoa" de la Universidad de Oviedo dentro de la investigación de Trabajo Fin de Máster de Samuel García Díaz titulada "Diseño y simulación de antenas metálicas para aplicaciones biológicas" [García Díaz, 2023]. Dicha investigación ha sido revisada y aprobada por el Comité ético del Principado de Asturias, así como la que se desarrolla en este trabajo, garantizando el cumplimiento de los estándares de ética y privacidad de los participantes.

Dentro del protocolo seguido en la investigación, se ha asignado un código a cada individuo para mantener su anonimato. Además, el acceso a los datos está restringido a miembros del equipo investigador.

Se cuenta con tres grupos compuestos por 100 individuos cada uno: un grupo de pacientes de cáncer de colon, otro de pacientes de cáncer de mama y un último de individuos de control de referencia. De cada participante se conocen adicionalmente su edad, sexo, diagnóstico y grado y etapa del cáncer. A continuación se presentan los rasgos más importantes de cada grupo.

1.2. Grupo de cáncer de mama

En este grupo sólo se cuenta con personas de sexo femenino con una edad de 61 ± 13 años en términos de media y desviación estándar.

En la Tabla 1.1 se puede observar un resumen de las características de los tumores diagnosticados, donde cabe destacar que de nuevo un caso era metastásico y, por otra parte, no en todos los individuos se pudieron medir la totalidad de las variables asociadas al diagnóstico.

Grado		**Tumor**		**Ganglios**	
Grado 1	31	PT1	54	N0	49
Grado 2	44	PT2	36	N1	20
Grado 3	23	PT3	9	N2	9
		PT4	1	N3	3

Tabla 1.1: Características de los casos de mama cáncer de colon en el estudio.

1.3. Grupo de cáncer de colon

De los 100 componentes de este grupo, 36 son de sexo femenino y 64 masculino, en la Tabla 1.2 se resumen las variables biológicas con las que contamos de los individuos y que podrían producir diferencias significativas en el espectro.

Sexo	**Nº de individuos**	**Edad (años: $\mu \pm \sigma$)**
Femenino	36	72 ± 11
Masculino	64	73 ± 9

Tabla 1.2: Características demográficas de los pacientes de cáncer de colon.

En la Tabla 1.3 se encuentran las principales características del diagnóstico realizado a los pacientes: el grado del cáncer, relacionado con su agresividad y la estadificación a través de las variables tumor primario (PT) que hace referencia al tamaño del tumor y ganglios linfáticos (N) que indica la cantidad de ganglios linfáticos tumorosos. Además, se sabe que sólo un caso es de metástasis. Cabe mencionar que en bastantes casos no fue posible asignar un grado al tumor y en algunos no fue posible medir el tumor primario o la afección a los ganglios.

Grado		Tumor		Ganglios	
Grado 1	5	PT1	11	N0	63
Grado 2	5	PT2	38	N1	23
Grado 3	5	PT3	48	N2	12
		PT4	1		

Tabla 1.3: Características de los casos de cáncer de colon en el estudio.

1.4. Grupo de control

Dentro de los individuos sanos de referencia, se cuenta con 42 personas de sexo femenino y 58 de sexo masculino. En la Tabla 1.4 se puede observar la edad de cada grupo en términos de media y desviación estándar.

Sexo	Número de individuos	Edad (años: $\mu \pm \sigma$)
Femenino	42	57 ± 4
Masculino	58	58 ± 5

Tabla 1.4: Características demográficas de los individuos de referencia.

Además, en la Figura 1.1, se pueden notar las diferencias en la distribución de la edad de los individuos de cada grupo de forma gráfica. Las edades del grupo de referencia están más concentradas, mientras que las de los grupos de cáncer está más dispersa. El grupo de cáncer de colon cuenta con individuos de mayor edad y el de cáncer de mama contiene los más jóvenes.

1.5. Espectroscopía infrarroja

Con el fin de reducir posibles errores experimentales, el procedimiento para la medidas de las muestras fue el siguiente.

- En primer lugar, se almacenaron las muestras homogeneizadas en un congelador a $-80°C$ en el departamento de biología molecular del Centro Científico - Técnológico Severo Ochoa.
- A continuación, las muestras fueron trasladadas a la sala de espectroscopía infrarroja, donde se obtuvieron los espectros infrarrojos de tres alicuotas mediante un espectrofotómetro FTIR Varian 620-IR, Figura 1.2(c). El espectro fue tomado en el rango de 900 a 3500 cm^{-1}, con 4 cm^{-1} de resolución espectral y 32 *scans* en cada espectro. Además, antes de la obtención de cada

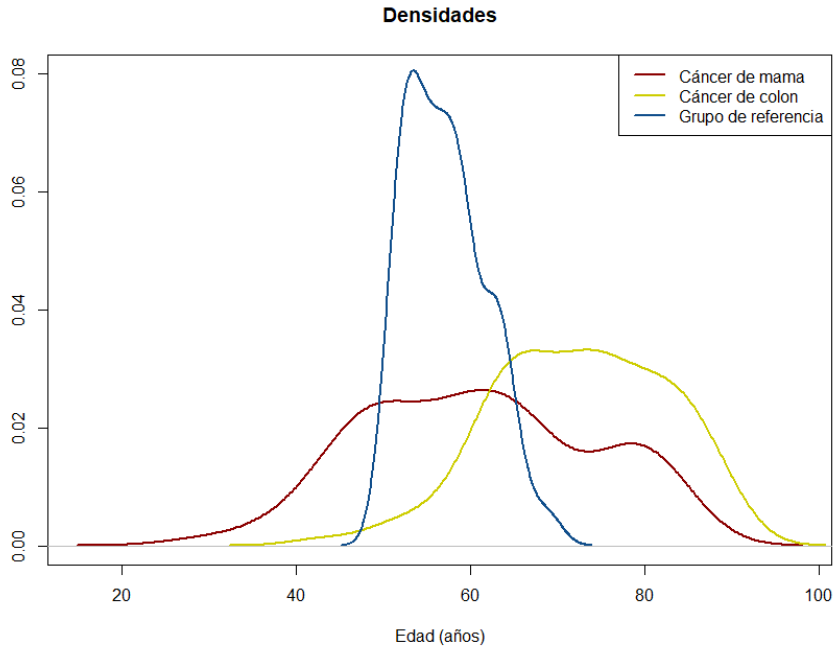
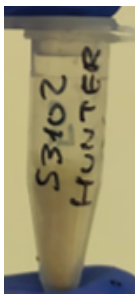
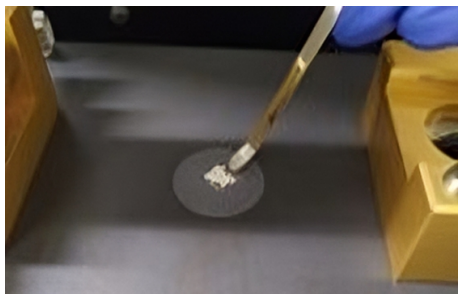


Figura 1.1: Representación de la densidad de la variable edad en cada grupo.

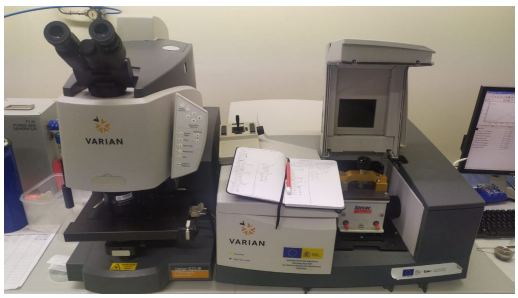
nuevo espectro, el cristal se limpió con agua destilada, acetona y alcohol y se midió un espectro del fondo para minimizar las variaciones del medio y otras posibles interferencias.



(a) *Muestra congelada*



(b) *Muestra homogeneizada*



(c) *Varian 620-IR*

Figura 1.2: Configuración experimental

En cuanto a la espectroscopía infrarroja por transformada de Fourier (espectroscopía FTIR), es una metodología analítica utilizada para estudiar la composición molecular de compuestos químicos. Se basa en el hecho de que las moléculas absorben una fracción de radiación electromagnética cuando incide sobre ellas. Esta absorción depende de la radiación incidente para cada enlace, creando el espectro que caracteriza los compuestos. Se utilizará la espectroscopía infrarroja en el infrarrojo medio, donde absorben la mayoría de enlaces moleculares, en concreto, los grupos funcionales. La espectroscopía infrarroja es una técnica que permite establecer una relación entre la radiación infrarroja y la materia,

lo que ofrece la posibilidad de identificar diferentes analitos biológicos y químicos, brindando un gran número de aplicaciones. Algunas de sus ventajas son la eficiencia en tiempo y costes y que no destruye la muestra.

Además, se utilizó la técnica ATR-FTIR, donde ATR proviene de las siglas en inglés de reflexión total atenuada, la que se ha convertido en la metodología estándar de medida de espectros con FTIR. Consiste en que el haz infrarrojo atraviesa un cristal de un cierto material ópticamente denso y con un alto índice de refracción en un determinado ángulo. Dicha reflectancia interna crea una onda evanescente que penetra la muestra que está en contacto con el cristal. Para aquellas regiones del espectro infrarrojo en el que la muestra absorba la onda evanescente que vuelve al cristal se verá atenuada por el detector situado en su opuesto. Dicha señal se utilizará como señal del interferograma al que se le aplicará la transformada de Fourier para pasar al dominio de número de onda. Una ventaja del uso del ATR y por lo que se ha extendido tanto su uso durante estos años es que permite analizar muestras de tamaño reducido y reduce los efectos de interferencia de agua o contaminantes superficiales.

Para analizar materiales biológicos, las regiones espectrales más importantes son la región del *fingerprint* compuesta por la huella dactilar (entre 600 y 1450 cm^{-1}) y la región de los picos de la amida I y la amida II (entre 1500 y 1700 cm^{-1}). La región de 2800 a 3500 cm^{-1} está asignada a vibraciones de estiramiento o tensión de enlaces como S-H, C-H, N-H y O-H, asociadas a lípidos, mientras que las regiones con números de onda más bajos suelen corresponder a vibraciones de flexión características de átomos de carbono en las moléculas [Baker et al., 2014]. Juntas, estas regiones constituyen una huella bioquímica de la estructura y función de las muestras a estudiar. Un espectro IR biológico típico con asignaciones moleculares se muestra en la Figura 1.3.

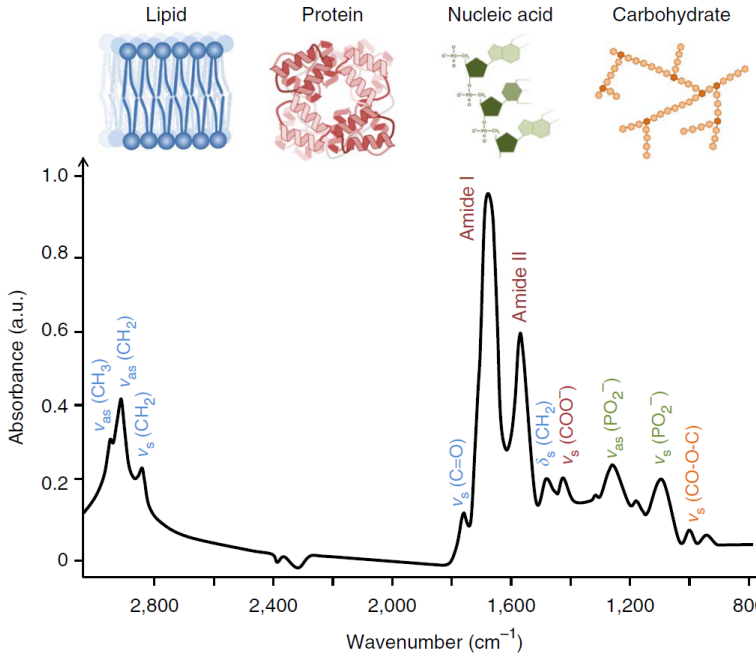


Figura 1.3: Espectro de una muestra de un carcinoma de mama humano donde se muestran las asignaciones de los picos de absorción entre 3000 cm^{-1} y 800 cm^{-1}. Figura tomada de [Baker et al., 2014].

Capítulo 2

Introducción a datos funcionales

En este capítulo se presentan los fundamentos teóricos necesarios para el análisis de la base de datos presentada en el capítulo anterior. Dado que está compuesta por espectros infrarrojos, de naturaleza intrínsecamente funcional, es necesario disponer de herramientas adecuadas para realizar un análisis correcto. Se comenzará definiendo el espacio funcional y sus medidas de localización. También se incluirán el análisis de componentes principales funcionales, FPCA, y la teoría ANOVA aplicada a datos funcionales. Finalmente, se describirán varios clasificadores: multivariantes, basados en estimación kernel, regresión binaria y profundidades. Pueden encontrarse más detalles en [Ferraty and Vieu, 2006, Ramsay and Silverman, 2005].

2.1. Motivación

La idea básica del análisis de datos funcionales es pensar en las trayectorias observadas como elementos individuales en vez de como una sucesiones de observaciones de variables individuales. El término funcional cuando se aplica a datos observados se refiere a la estructura intrínseca de los datos, más que a su forma explícita. En la práctica, los datos funcionales suelen ser observados y registrados de manera discreta en pares, en este caso, $(\lambda_j, \mathcal{X}(\lambda_j))$ donde $\mathcal{X}(\lambda_j)$ es la medida de la función en el número de onda λ_j, probablemente perturbado por ruido de medida.

Lo que hace las observaciones discretas funcionales es la suposición de la existencia de una función $\mathcal{X}(\lambda)$ que da lugar a los datos observados. Además, normalmente se quiere asumir que la función subyacente es suave, de manera que un par de valores del dato consecutivos tengan un cierto nivel de correlación y no difieran mucho entre ellos. Si no se diera el caso de este suavizado no habría mucha diferencia entre tratar los datos como funcionales o multivariantes.

En este contexto, la elección del espacio y la métrica en los que se trabaja adquiere una gran importancia. Estos elementos determinan cómo se miden las similitudes y diferencias entre las funciones, lo cual influye directamente en la precisión y la eficacia de los métodos estadísticos aplicados. Además, dependiendo del espacio en el que se trabaje, se tendrán disponibles unas herramientas estadísticas o no. Por ejemplo, el caso ideal de encontrarse en el espacio de funciones de Lebesgue $\mathcal{L}_2$, asegura que operaciones matemáticas como la diferenciación o la proyección estén bien definidas y garantiza propiedades de convergencia de sucesiones de funciones.

Una correcta elección del espacio funcional y la métrica no solo proporciona una base teórica sólida,

sino que también mejora la interpretabilidad y la robustez de los resultados obtenidos. Esto es esencial para garantizar la estabilidad y la fiabilidad de las estimaciones y modelos construidos a partir de los datos funcionales, haciendo que el análisis sea más adecuado para las características intrínsecas de los datos estudiados.

2.2. Primeras definiciones

Antes de comenzar el análisis de la base de datos, se definirán algunos conceptos necesarios para entender los datos funcionales y establecer las bases estadísticas y matemáticas necesarias para el análisis posterior de la base de datos.

Definición 2.1. *Un espacio* ***métrico*** *(****semimétrico*****) es un par* (E, d) *donde* E *es un conjunto no vacío y* $d : E \times E \to \mathbb{R}$ *es una función métrica, esto es que cumple: no negatividad, simetría, desigualdad triangular e identidad indiscernible*.*

Además, será útil que el espacio métrico cumpla algunas propiedades que garanticen que las técnicas de estimación y predicción sean válidas, como ser cerrado o ser completo. También se puede suponer estructura de espacio vectorial, es decir que los elementos puedan sumarse o multiplicarse por escalares sin salirse del espacio.

Definición 2.2. *Una variable aleatoria* $\mathcal{X}$ *se llama* ***variable funcional*** *(v.f.) si toma valores en un espacio métrico o semi-métrico completo* $\mathcal{F}$*, espacio funcional. Una observación* $\mathcal{X}_i$ *de* $\mathcal{X}$ *se dice* ***dato funcional****.*

Algunos ejemplos típicos de espacios funcionales son los espacios $\mathcal{L}_p$,

$$\mathcal{L}_p[\mathcal{S}, \mu] = \{f : \mathcal{S} \to \mathbb{R} \text{ de manera que } \int |f|^p < \infty\},$$

con $(\mathcal{S}, \mu)$ un espacio de medida y $p \in [1, \infty)$.

Definición 2.3. *Un* ***conjunto de datos funcional*** $\{\mathcal{X}_i\}_{i=1}^n$ *es la observación de n variables funcionales idénticamente distribuidas según* $\mathcal{X}$*.*

Se debe elegir el espacio funcional que mejor describa la situación estadística y tener en cuenta sus características. Por ejemplo, en un espacio métrico se cuenta tan solo con las distancias entre las curvas, mientras que en un espacio de Bannach el uso de la norma facilita realizar operaciones algebraicas y numéricas, mejorando la escalabilidad. Por su parte, en un espacio de Hilbert se cuenta con un producto interior que permite realizar descomposiciones espectrales y generalizar conceptos de geometría euclidiana.

Una clase importante de espacios funcionales son los espacios de Lebesgue, en particular el espacio $\mathcal{L}_2$ por sus propiedades estructurales que serán aprovechadas en este trabajo.

Definición 2.4. *Sea el conjunto de funciones definidas en el intervalo* $\mathcal{S} = [a, b]$ *cuya p-ésima potencia es integrable, es decir:*

$$\mathcal{L}_p := \left\{ f : \mathcal{S} \to \mathbb{R} \text{ tal que que } \int |f|^p < \infty \} \right\}.$$

Este espacio, con la métrica $d_p = \left(\int_{\mathcal{S}} |f(t) - g(t)|^p dt\right)^{1/p}$*, es un espacio métrico denominado* ***espacio de Lebesgue****. Más aún, cada espacio* $\mathcal{L}_p$ *es un espacio normado con la norma* $\|f\|_p := \left(\int_{\mathcal{S}} |f(t)^p dt|\right)^{1/p}$*. Además, el espacio* $\mathcal{L}_2$ *es un espacio de Hilbert con el producto interior* $\langle f, g\rangle := \int_{\mathcal{S}} f(t)g(t)dt$*.*

A partir de ahora, se considerará que los datos se encuentran en un espacio de Hilbert, que tiene ciertas propiedades interesantes, en particular dispone de un sistema base. Esto es un conjunto de funciones ϕ_k conocidas que permite representar cualquier función mediante una combinación lineal de elementos de la base

$$\mathcal{X}(t) = \sum_{k\in\mathbb{N}} c_k \phi_k(t).$$

Hay dos clases de bases: las fijas, que no dependen de los datos, como por ejemplo bases de Fourier, wavelets o de B-Splines y las bases dependientes de los datos, como la de componentes principales y la de mínimos cuadrados. A lo largo de este trabajo se utilizó la representación en una base de componentes principales cuando fue necesaria.

2.2.1. Representación en una base de Componentes Principales Funcional

La representación de los datos en una base de componentes principales busca encontrar un conjunto de K funciones ortonormales ν_k que permitan una expansión de cada curva $\mathcal{X}_i$ [Ramsay and Silverman, 2005]. Estas funciones provienen de la descomposición espectral de la matriz de varianzas-covarianzas. Así, los *scores* de las componentes principales correspondientes a una componente ν se definen como $f_i = \langle \nu, \mathcal{X}_i\rangle = \int \nu(s)\mathcal{X}_i(s)\, ds$. La primera función de peso $\nu_1(s)$ se elige para maximizar $N^{-1}\sum_i f_{i1}^2$, sujeto a la restricción $\int \nu_1(s)^2\, ds = 1$. Para esto, es esencial contar con el producto interior funcional $\langle \beta, \mathcal{X}\rangle = \int \nu(s)\mathcal{X}(s)\, ds$, que reemplaza la suma discreta del contexto multivariante por una integral. De esta manera, la representación de los datos viene dada por

$$\mathcal{X}_i \approx \mu + \sum_{j=1}^{K} \langle \mathcal{X}_i, \nu_j\rangle \nu_j = \mu + \sum_{j=1}^{K} c_{ij}\nu_j,$$

con μ la media de la variable aleatoria funcional, que se abordará en la siguiente sección.

El criterio de ajuste de la base se mide mediante el error cuadrático integrado $\|\mathcal{X}_i - \hat{\mathcal{X}}_i\|^2 = \int [\mathcal{X}(s) - \hat{\mathcal{X}}(s)]^2\, ds$. La base de componentes principales minimiza el error total $PCASSE = \sum_{i=1}^{N} \|\mathcal{X}_i - \hat{\mathcal{X}}_i\|^2$. Esta elección de funciones base maximiza las componentes de varianza, siendo denominadas funciones ortonormales empíricas.

Una vez definido el marco de trabajo se procede a introducir los métodos estadísticos de los que se hizo uso a lo largo de este trabajo junto con sus particularidades al estar en el contexto de datos funcionales.

2.3. Medidas de localización para datos funcionales

Definición 2.5. *Sea* $\mathcal{X}$ *una variable aleatoria funcional que toma valores en el espacio métrico* $\mathcal{F}$ *se define la* ***media*** *de* $\mathcal{X}$ *como:*

$$\arg\min_{a\in\mathcal{F}} \sum_{\mathcal{X}\in\mathcal{S}} d(\mathcal{X}, a)^2,$$

que se corresponde con el centro de gravedad. De la misma manera, sea $\{\mathcal{X}_i\}_{i=1}^n$ *un conjunto de datos funcionales con* $\mathcal{X}_i \overset{iid}{\sim} \mathcal{X}$ *se define la* ***media muestral*** *como:*

$$\arg\min_{a\in\mathcal{X}_n} \sum_{i=1}^{n} d(\mathcal{X}_i, a)^2.$$

En un espacio de Hilbert se tiene un indicio de estructura de $\bar{\mathcal{X}}$ al contar con una base para representarlo $\bar{\mathcal{X}} = \sum_{j\in\mathbb{N}} c_j\phi_j$. Además, el problema de minimización de la distancia al cuadrado en la práctica, representada por el producto interior, pasa por minimizar una forma cuadrática con matriz definida positiva, por lo que el mínimo se obtiene para el centro de la elipse $\bar{c}_j = \frac{1}{n}\sum_{i=1}^n c_{ij}$

A pesar de las buenas propiedades de la media es sabido que carece de robustez frente a datos atípicos, por lo que se presenta como alternativa robusta de medida de localización la mediana. Una manera de definir la mediana de una variable aleatoria real X es como la solución bajo existencia y unicidad del problema de minimización $\text{ínf}_{x\in\mathbb{R}} E(|x-X|)$. Se puede extender esta idea al caso funcional reemplazando $\mathbb{R}$ por $\mathcal{F}$ y $|x|$ por d.

Definición 2.6. *Sea* $\mathcal{X}$ *una variable aleatoria funcional que toma valores en el espacio métrico* $\mathcal{F}$ *se define la* ***mediana*** *de* $\mathcal{X}$ *como:*

$$\arg\min_{a\in\mathcal{F}} \sum_{\mathcal{X}\in\mathcal{S}} d(\mathcal{X}, a).$$

De la misma manera se define su análogo muestral, sea $\{\mathcal{X}_i\}_{i=1}^n$ *un conjunto de datos con* $\mathcal{X}_i \overset{iid}{\sim} \mathcal{S}$*, la* ***mediana muestral*** *se corresponde con la solución a*

$$\arg\min_{a\in\mathcal{X}_n} \sum_{i=1}^{n} d(\mathcal{X}_i, a).$$

Mientras que la media es la medida que minimiza el error medio cuadrático, la mediana que minimiza el error medio absoluto, y no se ve tan afectada frente a atípicos.

Otra herramienta estadística útil en el análisis de datos funcionales es la profundidad estadística, que es una medida que cuantifica la distancia de un punto con respecto al resto de los datos en una muestra. Cuanto mayor sea la profundidad de un punto, este se considera más centrado o representativo en relación a la muestra. La profundidad estadística permite introducir un orden en la muestra, permitiendo definir estimadores robustos de medidas de tendencia central como el elemento de la muestra más profundo $\arg\max_{a\in\mathcal{X}_n} D(\mathcal{X}_i, a)$ y también facilita la identificación de atípicos, siguiendo metodologías los métodos presentados en [Febrero et al., 2007]. Además, los conceptos de profundidad de datos pueden utilizarse en clasificación, asignando un nuevo dato según su profundidad relativa en las diferentes muestras de entrenamiento

Las profundidades que se han utilizado a lo largo de este trabajo son las siguientes:

Definición 2.7. *Sean* $\{\mathcal{X}_i(t)\}_{i=1}^n$ *realizaciones independientes e idénticamente distribuidas de una variable aleatoria funcional con dominio* $\mathcal{T} = [a, b]$*. Sea* D *una profundidad en* $\mathbb{R}$*, para cada* $t_0 \in \mathcal{T}$*, considerando* $z_i(t_0) := D(\mathcal{X}_i(t_0))$ *como la profundidad univariante del dato* i *en* t_0 *con respecto a* $\{\mathcal{X}_i(t_0)\}_{i=1}^n$ *se define la* ***profundidad de Fraiman-Muniz*** *[Fraiman and Muniz, 2001], como:*

$$FMD(\mathcal{X}_i) = \int_{\mathcal{T}} z_i(t)dt.$$

Por lo que la profundidad de Fraiman-Muniz, a partir de ahora FM, puede verse como el promedio de una profundidad univariante a lo largo de $\mathcal{T}$*.*

Definición 2.8. *Sean $\{\mathcal{X}_i(t)\}_{i=1}^n$ realizaciones independientes e idénticamente distribuidas de una variable aleatoria funcional. Sea $K : \mathbb{R}^+ \to \mathbb{R}^+$ una función kernel asimétrica y h el parámetro de suavizado. Se define la* ***profundidad modal*** *como:*

$$MD(\mathcal{X}_i) := \sum_{i=1}^{n} K\left(\frac{d(\mathcal{X}_i, \mathcal{X}_j)}{h}\right).$$

Esta profundidad es una medida de cuántos puntos se encuentran en la vecindad. Se asemeja al estimador kernel no paramétrico de densidad con la diferencia de que el parámetro ventana no va necesariamente a cero cuando n aumenta. Aquí el objetivo es asignar rangos que no cambien para anchos de banda suficientemente grandes.

Definición 2.9. *Sean $\{\mathcal{X}_i(t)\}_{i=1}^n$ realizaciones independientes e idénticamente distribuidas de una variables aleatoria funcional. Sea $h \in \mathcal{H}$ una realización del proceso de dirección independiente $\mathcal{H}$ y $P_i^h = \langle h, \mathcal{X}_i\rangle \in \mathbb{R}$ la proyección de $\mathcal{X}_i$ a lo largo de la dirección h. Se define la* ***profundidad de proyecciones aleatorias*** *como:*

$$RPD(\mathcal{X}_i, h) := D(P_i^h),$$

siendo D una medida de profundidad univariante.

En la práctica, es importante tener asociada una medida de incertidumbre sobre la estimación realizada. En el caso de las medidas de tendencia central se calcularán bandas de confianza mediante remuestreo uniforme siguiendo el procedimiento expuesto en [Cuevas et al., 2006]. Sea $\mathcal{X}_n = \{\mathcal{X}_i\}_{i=1}^n$ la muestra y $\hat{\theta}(\mathcal{X}_n)$ el estimador de localización de $\theta(\mathcal{X})$. La banda de confianza $1-\alpha$ de $\hat{\theta}(\mathcal{X}_n)$ se define a partir del cuantil $q_{1-\alpha}$ de las distancias $d(\hat{\theta}(\mathcal{X}_n), \hat{\theta}(\mathcal{X}_n^*))$ obtenido mediante remuestreo.

El enfoque no paramétrico del bootstrap resulta especialmente útil en el análisis de datos funcionales, donde generalmente no se disponen de distribuciones conocidas. Cabe destacar que no se corresponden con bandas como tal salvo el caso de la distancia del supremo.

2.4. Análisis de las Componentes Principales Funcionales

La representación de datos funcionales en una base puede usarse también como herramienta de análisis exploratorio, especialmente para las componentes principales. Dado que la base proviene de la descomposición espectral de la matriz de varianzas-covarianzas, como en el caso multivariante, sirve para analizar la esctructura de variabilidad de las curvas. Así, la aproximación de un dato funcional según su base de componentes principales viene dado en la práctica por

$$\mathcal{X}_i \approx \bar{\mathcal{X}} + \sum_{j=1}^{K} \langle \mathcal{X}_i, \hat{\nu}_j\rangle \hat{\nu}_j = \bar{\mathcal{X}} + \sum_{j=1}^{K} \hat{c}_{ij}\hat{\nu}_j,$$

donde $\bar{\mathcal{X}}$ representa la media de los datos, lo que es común a todos ellos y que se extrae del análisis, por lo que la FPCA captura directamente los principales modos de variación alrededor de ella en las funciones principales ν_j.

2.5. ANOVA Funcional

Un problema estadístico natural es el de decidir la existencia o no de diferencias en un proceso de interés al cambiar algunas condiciones que lo pueden afectar. Se pueden estudiar distintas configuraciones del problema, como se recoge en [Cuesta-Albertos and Febrero-Bande, 2010], sin embargo en esta memoria se hará uso de dos modelos. En primer lugar, se comienza estudiando el modelo unifactorial, en el que se examinan las diferencias en media de una variable de respuesta entre dos o más grupos definidos por un solo factor. De manera formal: dada una muestra de datos funcionales clasificados por una variable que agrupa, es decir $\{\mathcal{X}_i, G_i\}_{i=1}^n \in \mathcal{F} \times \mathbb{G} = \{1, \ldots, G\}$ donde G es una variable discreta que indica el grupo al que pertenece cada observación. El objetivo es contrastar la hipótesis nula

$$\mathcal{H}_0 : \bar{\mathcal{X}}_1 = \bar{\mathcal{X}}_2 = \ldots = \bar{\mathcal{X}}_g,$$

contra la alternativa

$$\mathcal{H}_1 : \exists k, j \text{ tal que } \bar{\mathcal{X}}_k \neq \bar{\mathcal{X}}_j.$$

Se pueden utilizar distintos métodos que manejan problemas ANOVA. Algunos de ellos sólo para este caso de diseño unifactorial, otros admiten diseños más complicados, pero son más costosos computacionalmente. También existen otros métodos basados en comparaciones univariantes puntuales, que pueden no funcionar bien incluso en casos sencillos. Otra limitación es que muchos de estos métodos están basados en hipótesis de normalidad.

En el caso unifactorial, se empleará el estadístico $V_n = \sum_{k<j} n_i \|\bar{\mathcal{X}}_k - \bar{\mathcal{X}}_j\|^2$. Este estadístico mide la variabilidad total entre los grupos al sumar las diferencias al cuadrado ponderadas por el tamaño de los grupos. La principal diferencia del caso funcional con el multivariante es que no se tiene una distribución de referencia con la que comparar el estadístico, por lo que se empleará la metodología bootstrap. Este enfoque permite evitar hipótesis de homocedasticidad, presente típicamente en los modelos ANOVA univariantes, a cambio de utilizar un test asintótico aproximado.

Se estudiará también el FANOVA multifactorial sin interacciones, para su formulación se empleará un modelo equivalente que asume la existencia de una función subyacente que describe la evolución típica del proceso considerado, asumiendo que los datos que se tienen han sido obtenidos añadiendo fluctuaciones aleatorias a dicha función típica. Sean $R, S \in \mathbb{N}$ y para cada $r = 1, \ldots, R$ y $s = 1, \ldots, S$ existen $\mathcal{X}_i^{r,s}(\lambda), i = 1, \ldots, n_{r,s} \in \mathbb{N}$ funciones aleatorias de manera que

$$\mathcal{X}_i^{r,s}(\lambda) = m(\lambda) + f^r(\lambda) + g^s(\lambda) + \epsilon_i^{r,s}(\lambda), \quad \lambda \in [a, b],$$

donde la función m es no aleatoria y describe la forma del proceso y f^r y g^s tienen en cuenta los efectos principales de los dos factores y su interacción. Es necesario asumir que para cada $\lambda \in [a, b], r_0 = 1, \ldots, R$, y $s_0 = 1, \ldots, S$,

$$\sum_r f^r(\lambda) = \sum_s g^s(\lambda) = 0.$$

Las trayectorias aleatorias $\epsilon_i^{r,s}$ se asumen independientes y centradas en media. Además, para cada r, s fijados, $\epsilon_i^{r,s}, i = 1, \ldots, n_{r,s}$ están idénticamente distribuidos.

Se busca testear la hipótesis nula de:

$$\begin{cases} \mathcal{H}_0^A : & f^1 = \ldots = f^R = 0, \quad \text{el primer factor no influye} \\ \mathcal{H}_0^B : & h^1 = \ldots = h^S = 0, \quad \text{el segundo factor no influye} \end{cases}$$

Para hacer frente a este diseño ANOVA se transformarán los datos funcionales en datos univariantes utilizando una proyección aleatoria. Se resolverá el problema ANOVA en esa situación simplificada obteniendo conclusiones para los datos funcionales a partir de la información aportada por varias proyecciones. Es importante destacar que de esta manera evitamos la necesidad de hipótesis de normalidad.

Sin embargo, este procedimiento tiene dos principales inconvenientes. Uno es la pérdida de información al pasar de un dato funcional a uno univariante y el otro la inestabilidad inherente al proceso de selección de una proyección aleatoria. En la práctica, para reducir el efecto de estas limitaciones, aunque teóricamente una proyección es suficiente, se propuso tomar $k > 1$ proyecciones aleatorias, contrastar la hipótesis nula en cada proyección y ajustar los p-valores. Para ajustar los p-valores se puede utilizar el procedimiento conservador de Bonferroni, el costoso computacionalmente bootstrap o una corrección que controla el *False Discovery Rate*, FDR, que es la proporción esperada de hipótesis erróneamente rechazadas.

2.6. Clasificadores para Datos Funcionales

El problema de la clasificación supervisada binaria se puede plantear, en general, de la siguiente manera: se cuenta con una muestra de entrenamiento totalmente clasificada por grupos $\{\mathcal{X}_i, G_i\}_{i=1}^n \in \mathcal{F} \times G = \{1, \ldots, G\}$ donde G es una variable discreta que indica el grupo al que pertenece cada observación. El objetivo es estimar las probabilidades *a posteriori* de que una nueva observación $\mathcal{X}$ pertenezca a cada grupo, es decir

$$p_g(\mathcal{X}) = \mathbb{P}(G = g|\mathcal{X} = \mathcal{X}) = \mathbb{E}\left[\mathbb{1}_{\{G=g\}}|\mathcal{X} = \mathcal{X}\right]. \tag{2.1}$$

La regla óptima de clasificación es asignar la nueva observación al grupo que maximiza la probabilidad *a posteriori*, esto es:

$$\hat{G}_{\mathcal{X}} = \arg\max_{g \in \mathbb{G}} \quad \hat{p}_g(\mathcal{X}).$$

Existen varias estrategias para afrontar el cálculo de la regla óptima de clasificación, estas son:

Multivariantes Una primera aproximación al problema es reciclar los clasificadores multivariantes utilizando una aproximación de las curvas a través de una representación en una base finita. La elección de la base en la que se representarán los datos es importante, ya que algunos modelos son sensibles a esta elección.

Estimación Kernel Por otra parte, dada la probabilidad en (2.1), también se podría abordar el problema de clasificación a través de la estimación de las probabilidades *a posteriori* aproximándolas a través de kernel como por ejemplo con el estimador de Nadaraya-Watson:

$$\hat{p}_{g,h}(X) = \frac{\sum_{i=1}^n \mathbb{1}_{G_i=g} K(h^{-1} d(X, \mathcal{X}_i))}{\sum_{i=1}^n K(h^{-1} d(X, \mathcal{X}_i))},$$

donde K es una función kernel asimétrica, d es una distancia y h es el parámetro ventana, que se elegirá por validación cruzada.

Regresión binaria Otra estrategia observando la esperanza de (2.1) sería considerar el problema de clasificación como un problema de regresión binaria aprovechando la representación en una base finita de los datos funcionales, por ejemplo mediante un modelo general lineal, GLM.

Profundidades Por último, otra estrategia para clasificar puede deducirse a partir del DD-plot. Este método basa su clasificación en la profundidad de una observación en cada grupo, asignándolo al grupo en el que tiene mayor profundidad. Esto podría verse como un sucedáneo de la densidad en el caso de los grupos funcionales, por lo que también podría ser una especie de análogo del cálculo de la probabilidad condicionada según la regla de Bayes. Algunas propiedades del DD-plot es que si ambos grupos son iguales, en el DD-plot aparecerá una línea diagonal y mientras más claramente separados estén los grupos el DD-plot se irá pareciendo más a una forma de L. Además, la dispersión del DD-plot reflejará la de los datos originales. Esta regla del clasificador según la máxima profundidad puede extenderse a clasificadores más sofisticados, combinando la reducción de dimensión al pasar al espacio de profundidades con clasificadores multivariantes como los expuestos anteriormente.

2.6.1. Métricas de evaluación

Una vez entrenados los clasificadores se evalúa su rendimiento, este es un aspecto fundamental ya que permite decidir como de bueno es el clasificador propuesto y como se compara con otros algoritmos para el mismo problema. Una buena manera de representar el rendimiento de un algoritmo de aprendizaje estadístico para el caso de un problema de clasificación binario es a través de la matriz de confusión, que tiene en cuenta las diversas probabilidades asociadas a las distintas decisiones que un clasificador puede tomar. Se define como:

Observación ╲ Predicción	Positivo (1)	Negativo (0)
Positivo (1)	VP	FN
Negativo (0)	FP	VN

En el caso a estudiar en este trabajo, negativo serán los individuos de control de referencia y positivo indica que es paciente de cáncer. Así, VP serán los pacientes clasificados como tal por el algoritmo (análogo para verdadero negativo, VN). Por otra parte, FN serán los pacientes predichos como sanos por el clasificador (análogo para falso positivo, FP). Luego, cuanto más pequeños sean los valores fuera de la diagonal de la matriz de confusión, mejor será el rendimiento del algoritmo.

A partir de la matriz de confusión se definen varias métricas:

- La exactitud, *Accuracy* en inglés, es la proporción de eventos bien clasificados por el problema, o bien:
$$Acc = \frac{VP + VN}{VP + FN + FP + VN}.$$
- La sensibilidad o TPR, es la tasa de verdaderos positivos, por lo que representa la capacidad de detectar positivos correctamente, esto es:
$$\text{TPR} = \frac{VP}{VP + FN}.$$

- La especificidad o TNR, es la tasa de verdaderos negativos, representando la capacidad del clasificador de detectar correctamente los casos negativos:

$$\text{TNR} = \frac{VN}{VN + FP}$$

- La precisión mide cómo de exactas son las predicciones positivas, también se la conoce como valor predictivo positivo, PPV, su fórmula es

$$\text{PPV} = \frac{VP}{VP + FP}. \tag{2.2}$$

- El valor F, $F1$, es útil en muestras no balanceadas, ya que es una combinación entre la precisión y la sensibilidad:

$$F1 = 2\,\frac{\text{Precisión} \cdot \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}} = \frac{2VP}{2VP + FN + FP} \tag{2.3}$$

Dependiendo del contexto será más útil una métrica u otra. Por ejemplo, la exactitud es la más extendida en problemas de aprendizaje estadístico, pero falla cuando el problema está desbalanceado, en tal caso es mejor utilizar la métrica $F1$; la sensibilidad y especificidad pueden ser importantes a la hora de querer minimizar falsos negativos o positivos. En el contexto de un problema médico, y más particularmente de cribado, dependiendo de la población objetivo será más útil estudiar una métrica u otra. Esto es, si se realiza un cribado en una población de riesgo se querrán minimizar los falsos negativos (y por tanto maximizar la sensibilidad), pero si por el contrario el cribado se realizase en una población joven se querrían minimizar los falsos positivos, esto equivale a valores altos de especificidad, que conllevan estrés y gasto de recursos médicos.

Cabe mencionar que los algoritmos se pueden optimizar según la métrica que a considerar, aunque por defecto optimizan el valor del *accuracy*. Así se mantendrá al realizar las tareas de clasificación, pero se calcularán también el resto de métricas para hacerse una visión más general del comportamiento de los algoritmos.

Capítulo 3

Análisis exploratorio de los datos

En este capítulo y con el fin de comprender la estructura subyacente de los datos y su variabilidad se realiza un análisis exploratorio de los mismos a través de herramientas específicas del análisis de datos funcionales, así como técnicas multivariantes adaptadas a este contexto. Estos métodos permiten representar medidas de tendencia central, detectar atípicos y realizar análisis de componentes principales, sentando así una base sólida para los posteriores análisis del trabajo.

Para el análisis de datos realizado en este y el siguiente capítulo se utilizó la librería `fda.usc` de `R` (Versión 2.1.0) [Febrero-Bande and Oviedo de la Fuente, 2012]. Esta librería proporciona funciones avanzadas para el análisis de datos, realizando análisis exploratorio y descriptivo de datos funcionales, explorando sus características más importantes, como las mediciones de profundidad o la detección de valores atípicos funcionales, entre otros. También se incluyen métodos para realizar ANOVA funcional unifactorial y multifactorial y para la clasificación supervisada.

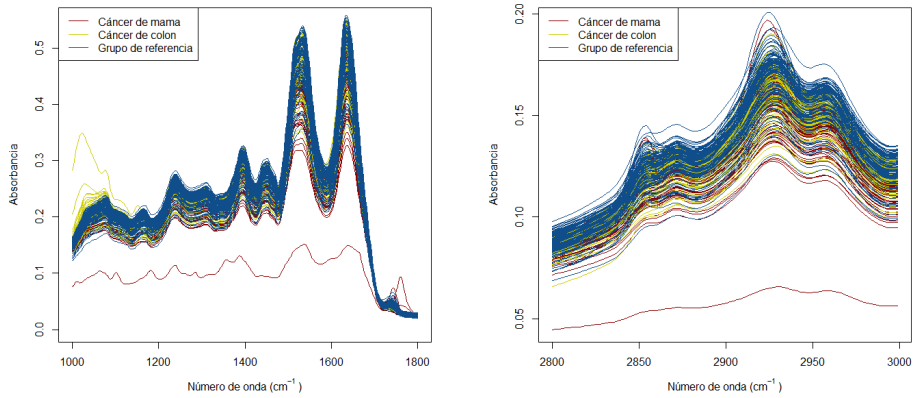
Una primera aproximación al análisis exploratorio puede ser la representación gráfica de las curvas en distintos colores según los grupos a los que pertenezcan. Sin embargo, esto puede ser engañoso ya que, como podemos comprobar en la Figura 3.1 las curvas de los grupos se superponen y es difícil detectar a simple vista patrones que los diferencien.

Cabe mencionar que se detecta a simple vista un claro atípico dentro del grupo de pacientes de cáncer de mama que se encuentra por debajo del resto de curvas en prácticamente la totalidad del espectro. Esta medida se eliminará y sería conveniente repetirla. Así, se pasa a la estimación de medidas de tendencia central para cada grupo.

3.1. Medidas de localización y dispersión

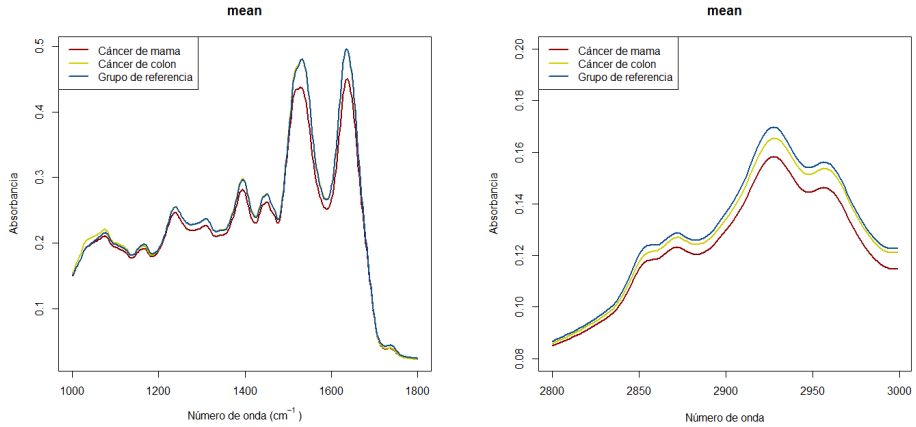
Esta sección se centrará en la identificación de patrones mediante medidas de tendencia central. En primer lugar, la media muestral, que es la medida de tendencia central más extendida y sencilla, Figura 3.2. Al estar en un marco de espacio de Hilbert y poder hacer uso de una base, se corresponderá con la media puntual. Además, se hará uso de la metodología bootstrap para estimar la incertidumbre asociada a dicha medida.

En la Figura 3.3 se pueden observar las medias de cada grupo junto con el 95 % de réplicas más cercanas a la media global calculadas por remuestreo. Así, se observa más variabilidad en la variable de la región de los lípidos y en la región del *fingerprint* en los picos de las amidas, entre 1500 y 1700 cm^{-1},



(a) *Región fingerprint* (b) *Región de lípidos*

Figura 3.1: Espectros infrarrojos: en rojo se representa el grupo de pacientes de cáncer de mama, en amarillo los de cáncer de colon y en azul el grupo de referencia.

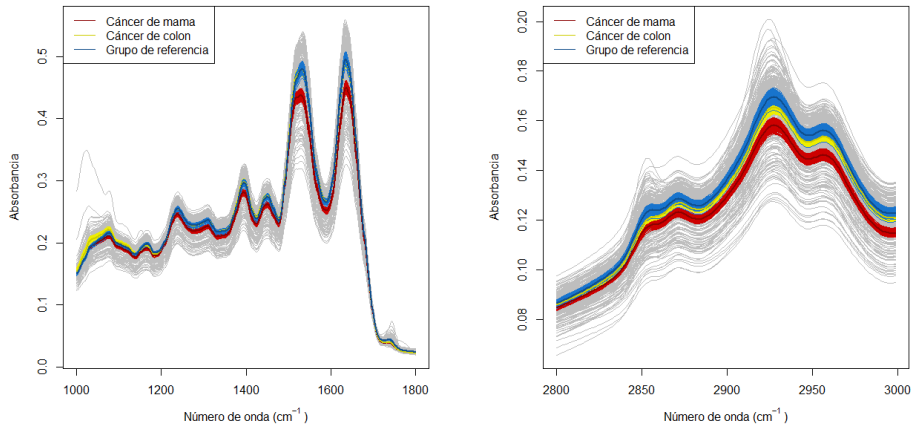


(a) *Región fingerprint* (b) *Región de lípidos*

Figura 3.2: Medias de los grupos, en rojo se representa el grupo de pacientes de cáncer de mama, en amarillo los de cáncer de colon y en azul el grupo de referencia.

que se refieren a las bandas de absorción características asociadas con los enlaces de las amidas, que son grupos funcionales comunes en proteínas. También se observa que, en media, el grupo de cáncer de mama se encuentra por debajo del resto en todo el espectro. Además, parece que en media la región de los lípidos sería la mejor para discriminar entre los todos los grupos, mientras que en casi toda la región del *fingerprint* se solapan los grupos de referencia y de cáncer de colon, salvo en la zona de ácidos nucleicos entre 1000 y 1200 cm^{-1}. Sin embargo, la diferencia en la zona de ácidos nucleicos puede ser debida a un dato influyente dentro del grupo de cáncer de colon que habría que estudiar si es atípico o no.

Dado el carácter poco robusto de la media, en ocasiones puede no ser la medida de tendencia central



(a) *Región fingerprint* (b) *Región de lípidos*

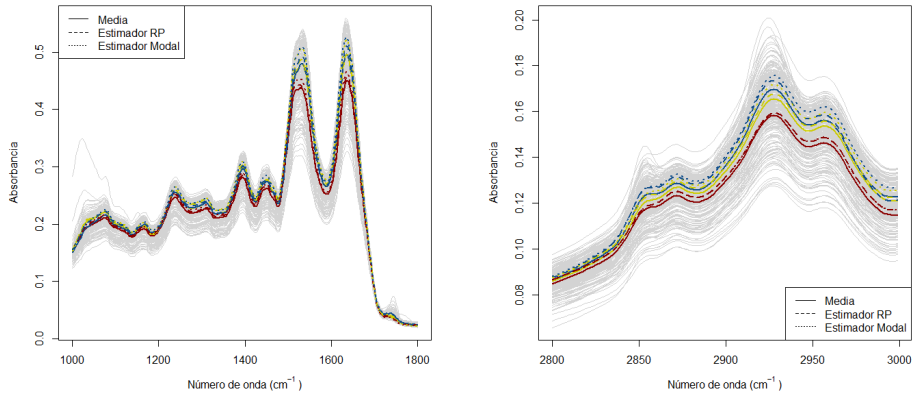
Figura 3.3: Medias de los grupos (en rojo el grupo de cáncer de mama, en amarillo el grupo de cáncer de colon y en azul el grupo de referencia) junto con una estimación de la incertidumbre asociada calculada por bootstrap. En rojo se representa el grupo de pacientes de cáncer de mama, en amarillo los de cáncer de colon y en azul el grupo de referencia.

más adecuada y conviene tener en cuenta versiones robustas como la mediana o los estimadores robustos de tendencia central basados en profundidades. En la Figura 3.4 se pueden observar la observación más profunda de cada grupo según la profundidad de proyecciones aleatorias, tomando 15 proyecciones, en línea discontinua. Además, en línea punteada aparece la versión según la profundidad modal y en línea continua la media muestral. En general, parece que la media está por debajo de los estimadores robustos, lo que se puede interpretar como un posible desplazamiento. Este efecto desaparecería en caso de diferenciar los espectros y en univariante, se correspondería con un desplazamiento hacia la izquierda.

Así como las medidas centrales se corresponden con el dato más profundo, aquellas observaciones menos profundas son susceptibles de ser atípicos, por lo que una manera de abordar la eliminación de atípicos es también mediante profundidades.

3.2. Detección de atípicos

Una manera de estudiar la existencia de atípicos es a través de remuestreo bootstrap. Se pueden hacer dos aproximaciones: una más agresiva en la que se recorta fijando una proporción de la muestra que se consideran atípicos y otra más suave ponderando pesos de profundidad. Sin embargo, por optimización de los tiempos de computación, se estudiaron los atípicos dentro de cada grupo mediante la metodología del suavizado y la profundidad FM. Estos resultados pueden observarse en la Tabla 3.1.



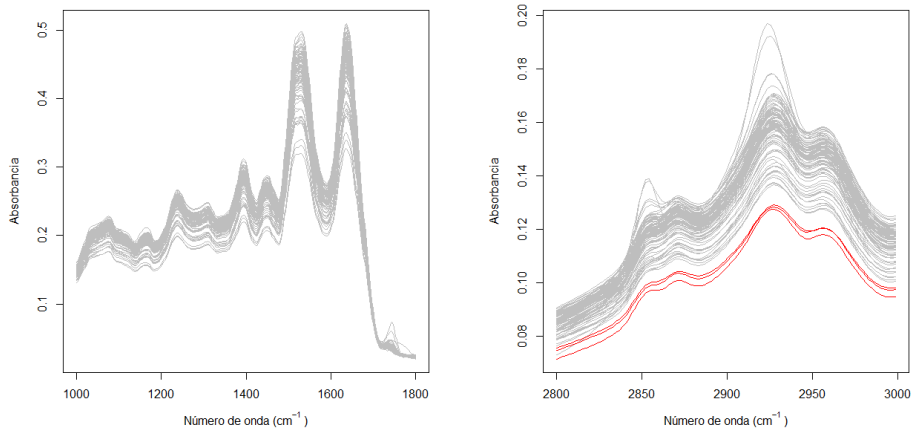
(a) *Región fingerprint* (b) *Región de lípidos*

Figura 3.4: Espectros de la muestra en gris, en rojo está representada la media del grupo de cáncer de mama, en amarillo la del grupo de pacientes de cáncer de colon y en azul la del grupo de referencia, con los estimadores robustos de localización según la profundidad de proyecciones aleatorias en línea discontinua y según la profundidad modal punteada.

Región	Grupo mama	Grupo colon	Grupo referencia
Fingerprint	–	$113, 198$	$229, 256$
Lípidos	$33, 43, 39$	113	$229, 256, 207$

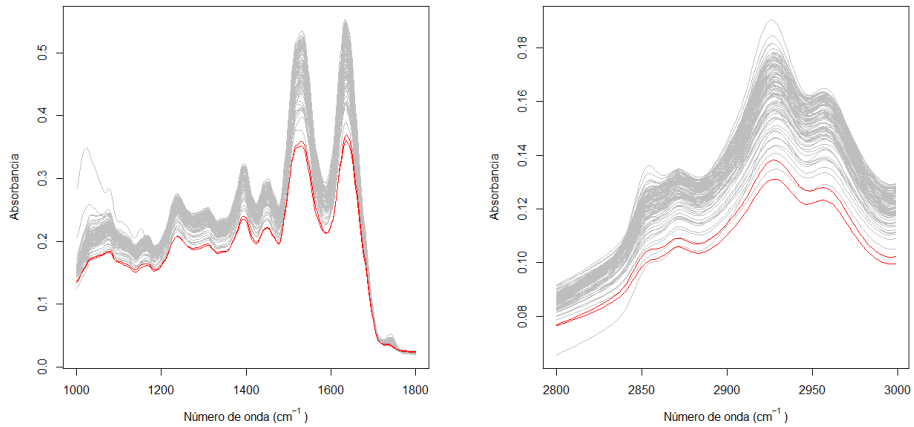
Tabla 3.1: Candidatos a atípicos estudiando cada grupo por separado utilizando la profundidad FM con un suavizado de $h = 0.1$.

Los espectros de los atípicos detectados en el grupo de cáncer de mama, de colon y de grupo de referencia pueden observarse en las Figuras 3.5, 3.6 y 3.7, respectivamente.



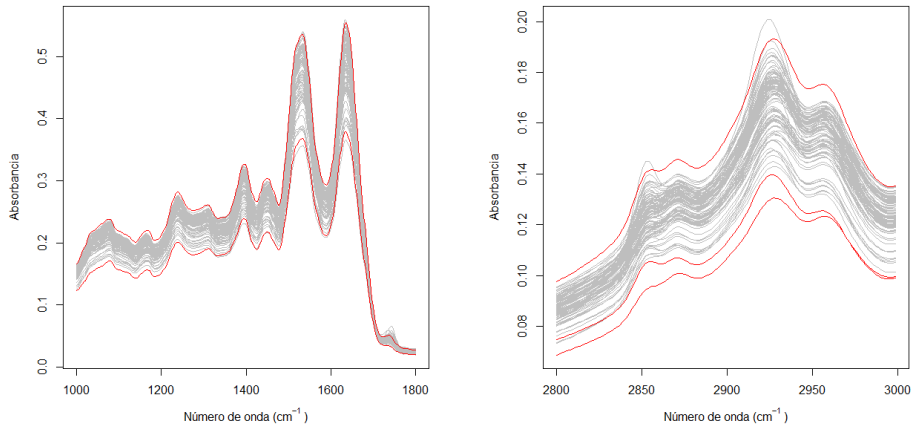
(a) *Región fingerprint* (b) *Región de lípidos*

Figura 3.5: Atípicos detectados en el grupo de cáncer de mama con la profundidad de FM con 1000 repeticiones bootstrap y suavizado $h = 0.10$.



(a) *Región fingerprint* (b) *Región de lípidos*

Figura 3.6: Atípicos detectados en el grupo de cáncer de colon con la profundidad de FM con 1000 repeticiones bootstrap y suavizado $h = 0.10$.

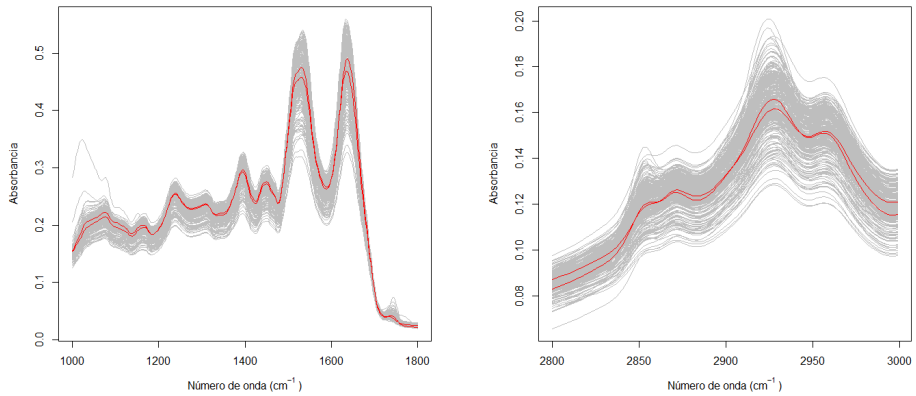


(a) *Región fingerprint* (b) *Región de lípidos*

Figura 3.7: Atípicos detectados en el grupo de referencia con la profundidad de FM con 1000 repeticiones bootstrap y suavizado $h = 0.10$.

Los algoritmos de detección de atípicos podrían llevar a una falsa detección, detectando valores que no se esperarían en sentido de tener una baja profundidad, sucedáneo de la densidad, pero sin ser atípicos. Por ello, se debe tener en cuenta el valor de la profundidad de los atípicos propuestos y el valor del corte obtenido por metodología bootstrap para decidir eliminar o no una observación. Así, no se eliminó ninguna muestra del grupo de cáncer de colon y tan solo se decide eliminar las observaciones 33 y 229 detectadas. Sus espectros se observan en al Figura 3.8 en rojo y a simple vista no se detecta ningún rasgo que indique que sean atípicos, reforzando la idea de que en análisis de datos

funcionales no todo es lo que parece. Por otra parte, pese a que las observaciones 39 y 43 también fueran detectadas en ambos casos, es posible que sean falsos atípicos ya que no se espera contar con tantos *outliers* en un mismo grupo de 100 individuos y son más profundas que la observación 33.



(a) *Región fingerprint* (b) *Región de lípidos*

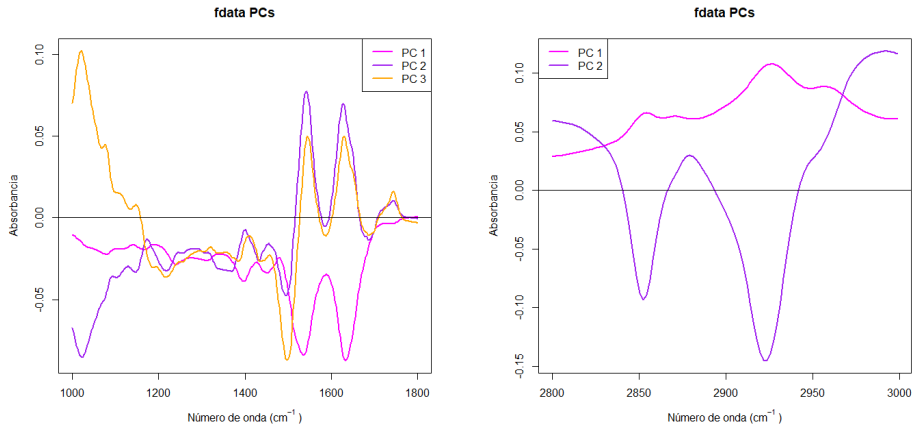
Figura 3.8: Atípicos eliminados del estudio teniendo en cuenta la profundidad FM.

3.3. FPCA

Se concluye este capítulo con un análisis de las componentes principales, esta técnica es clave en el ámbito del análisis de datos funcionales ya que proporciona un análisis exploratorio de los datos fácilmente interpretable y resume la variabilidad del conjunto. Además, es la base de otros modelos estadísticos como la regresión por componentes principales.

Se aplicará al análisis de componentes principales funcional a cada variable de la muestra eligiendo el número de componentes principales en cada región según el porcentaje de variabilidad que expliquen, de manera que en conjunto quede más del 97.5 % de la variabilidad total explicada. En la región del *fingerprint* dicho umbral se alcanza tomando 3 componentes principales, por lo que serán las componentes principales que se consideren. De estas tres componentes principales, la primera explica un 91.51 % de variabilidad total, la segunda un 4.48 % y la tercera un 2.34 %, sumando 98.33 % de variabilidad total explicada. Por otra parte, en la región de los lípidos con 2 componentes principales se explica un 97.82 % de la variabilidad total, del que un 92.12 % se corresponde a la primera componente y un 5.70 % a la segunda. Las componentes principales en cada región pueden observarse en la Figura 3.9. La primera componente principal en la región del fingerprint nunca corta al cero, que es la media, y se encuentra siempre por debajo. Por lo tanto, el primer patrón de variación es si la curva esta por encima o debajo (*score* negativo o positivo) de la media a lo largo de toda la curva. En cuanto a las otras dos componentes principales en esta región parecen bastante similares y opuestas salvo por un pico sobre los 1500 cm^{-1}. Además, ambas están afectadas por la observación del grupo de cáncer de colon que se encuentra por encima en la región de 1000 cm^{-1}. Esto se ve también en los diagramas de dispersión de los *scores* correspondientes a esta región en la Figura 3.10, donde se encuentra alejada de la nube de puntos principal. Se podría considerar eliminar esta observación del estudio como un posible

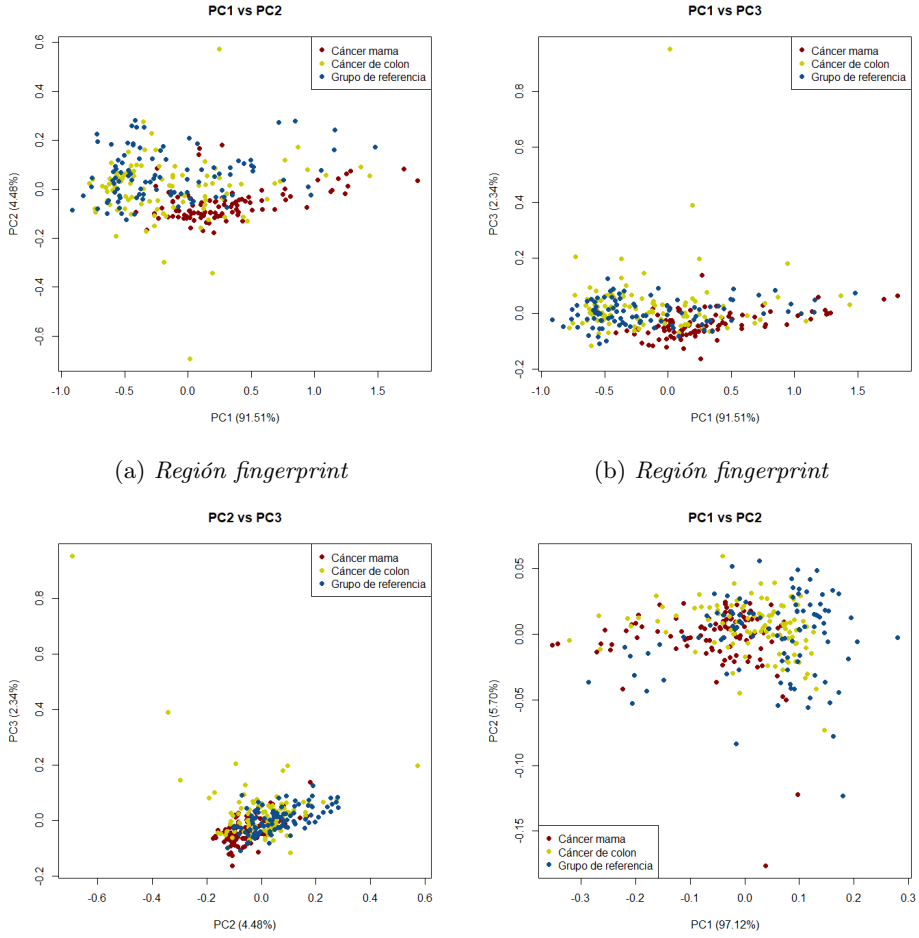
atípico. Por otra parte, en cuanto a las componentes principales en la región de los lípidos, en la primera el patrón de nuevo nunca corta el 0, por lo que se tendría la misma interpretación que en el caso de la región de fingerprint. El patrón de la segunda componente es más difícil de interpretar porque oscila entorno a la media. Cabe destacar que la primera componente principal tiene un pico de bastante altura en torno a los 2850 cm^{-1} relacionado con los ácidos grasos al corresponderse con el modo de vibración simétrica del CH_2. Es de especial interés que el grupo de pacientes de cáncer de mama se distinga del resto por estar relacionada con esta componente principal porque ya en [Blat et al., 2019] y en [Kepesidis et al., 2021] se relacionó esta región de los lípidos con el avance del cáncer de mama.



(a) *Región fingerprint* (b) *Región de lípidos*

Figura 3.9: Componentes principales en cada región espectral. Se tomaron 3 en el caso del fingerprint y 2 en la región de los lípidos, que es menos extensa.

En cuanto a las proyecciones en las componentes principales obtenidas, Figura 3.10, parece que hay bastante solape entre los grupos en general. Podría decirse que el grupo de pacientes de cáncer de mama se encuentra por encima del cero en la primera componente principal en la región del fingerprint (a la derecha del gráfico), por lo que puede tener una relación positiva con esta componente que representa la máxima variabilidad de los datos, esto podría deberse a que en el grupo de cáncer de mama los espectros se suelen situar por debajo de la media global. En esta región los grupos de referencia y de colon se encuentran bastante superpuestos y diferenciados del de cáncer de mama. Sin embargo, en el diagrama de dispersión de los *scores* de la región de lípidos se encuentran separados los grupos de colon y referencia cada uno a un lado de la primera componente principal, por lo que cada uno estará relacionado de manera positiva o negativa con este patrón de mayor variabilidad.



(a) *Región fingerprint* (b) *Región fingerprint*

(c) *Región fingerprint* (d) *Región de lípidos*

Figura 3.10: Proyecciones en las distintas componentes principales.

Capítulo 4

Modelos de clasificación

En este capítulo se abordarán las tareas de clasificación binaria grupo de pacientes de cáncer de mama y grupo de cáncer de contra grupo de referencia. En primer lugar y con el fin de definir los clasificadores de la mejor manera, se estudiará si el sexo, la edad y el grupo son factores en el problema contrastando la hipótesis nula de que las medias en cada grupo de las variables con iguales $\mathcal{H}_0 : \mu_1 = \mu_2$.

4.1. ANOVA

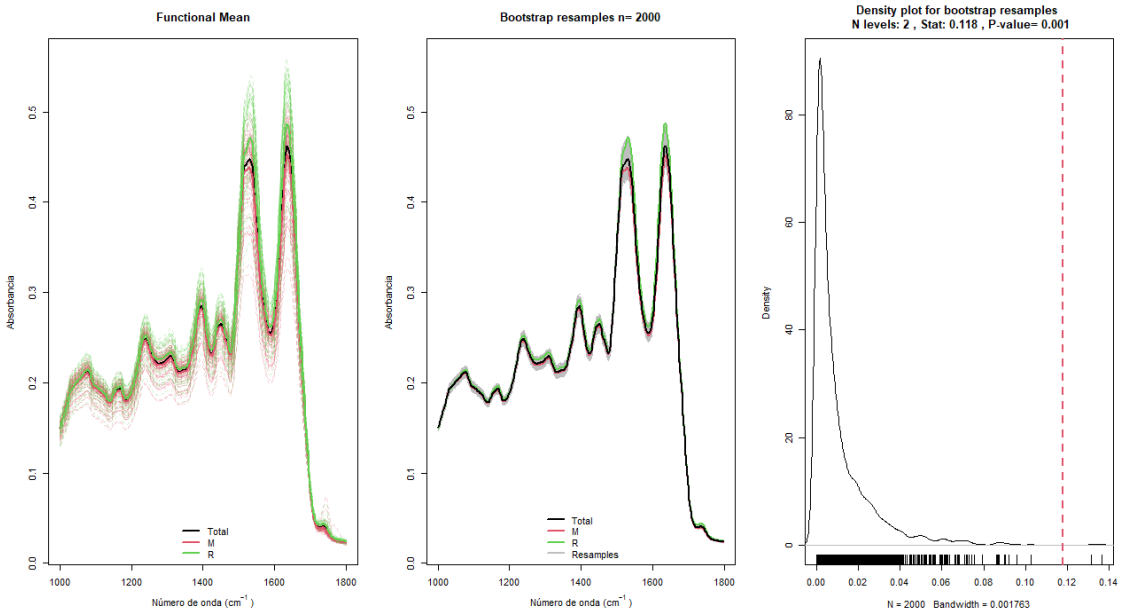
Con el fin de determinar la influencia de distintos factores en las variables funcionales de estudio se realizará análisis de la varianza funcional, primero unifactorialmente y después multifactorial.

4.1.1. Factor grupo

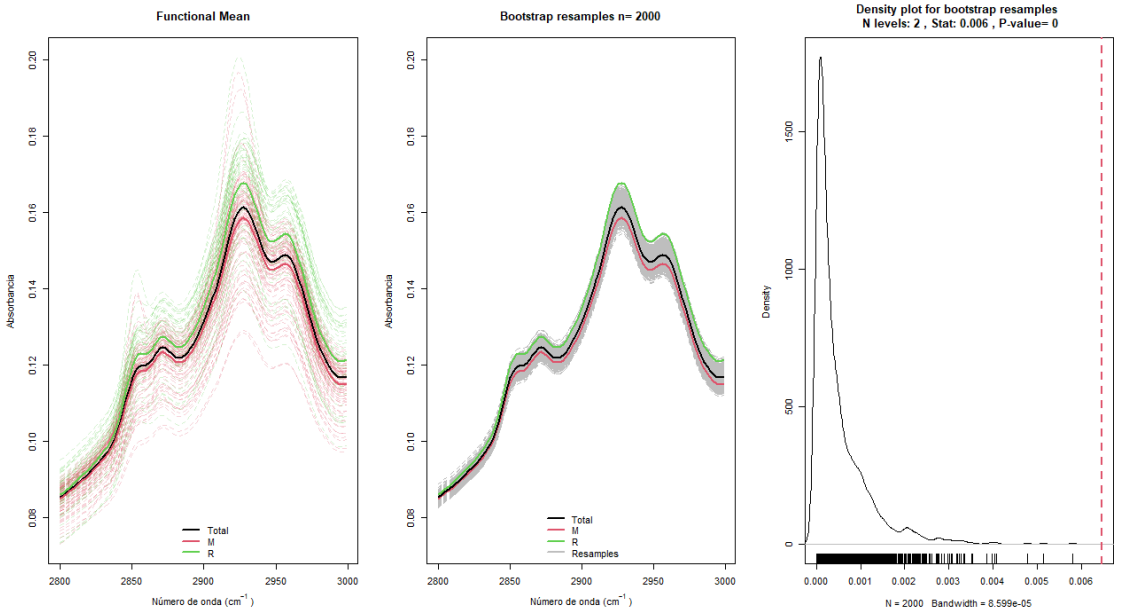
Se comienza estudiando el factor de grupo, en caso de que se detecten diferencias significativas entre grupos se puede intuir que el problema de clasificación resultará más sencillo a los clasificadores.

En el caso del grupo de cáncer de mama, para evitar el efecto de variables de confusión, sólo se toman las mujeres del grupo de control como referencia. Sin embargo, de esta manera quedan muy desbalanceados los grupos y podría afectar al resultado. Por ello, se procedió primero realizando el ANOVA con los grupos desbalanceados y a continuación se crearon remuestras de las de control hasta balancear el problema. Se tomó el mayor p-valor, esto es, el más desfavorable a lo que se querría. En ambas variables, *fingerprint* y lípidos, esto ocurrió en el caso original desbalanceado. Así, se concluye que se detectan cambios significativos entre las medias de los dos grupos, Figura 4.1.

Para el contraste de igualdad de medias entre el grupo de cáncer de colon y de referencia se observaron diferencias significativas en el variable de la región de los lípidos con un p-valor de 0.01, menor a los niveles de significación usuales, por lo que puede ser que esta variable sea más informativa a la hora de realizar la clasificación, Figura 4.2.

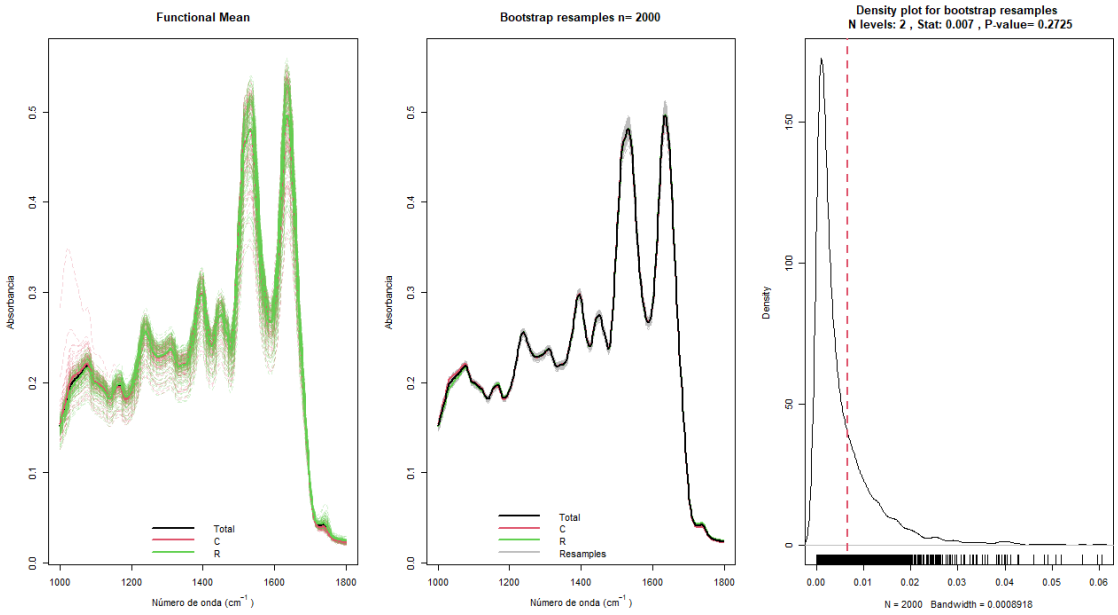


(a) *Región fingerprint*

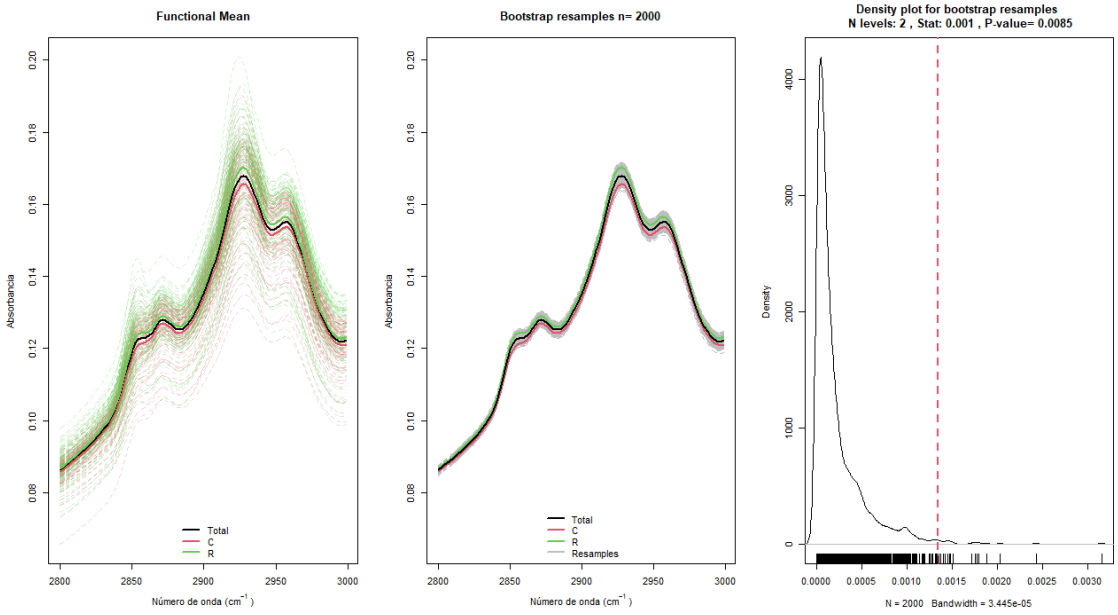


(b) *Región lípidos*

Figura 4.1: ANOVA unifactorial para el contraste de igualdad de medias entre el grupo de referencia y el de pacientes de cáncer de mama con $B = 2000$ réplicas bootstrap.



(a) *Región fingerprint*



(b) *Región lípidos*

Figura 4.2: ANOVA unifactorial para el contraste de igualdad de medias entre el grupo de referencia y el de pacientes de cáncer de colon con $B = 2000$ réplicas bootstrap.

En la Tabla 4.1 puede observarse un resumen de los p-valores obtenidos en los contrastes realizados de este apartado.

Grupo	**p-valor fingerprint**	**p-valor lípidos**
Cáncer de mama y grupo de referencia	0.001	0
Cáncer de colon y grupo de referencia	0.27	0.01

Tabla 4.1: Resumen de los p-valores obtenidos para los contrastes de igualdad de medias del factor grupo.

4.1.2. Factor edad

Como se puede comprobar en la Figura 1.1, las edades en cada grupo están distribuidas de manera muy diferente. Debido a este sesgo en la edad por grupos, se aceptará que esta variable es un posible factor de confusión en el problema. En el caso del grupo de cáncer de colon, al entrenar clasificadores una regla de clasificación podría ser la edad.

4.1.3. Factor sexo

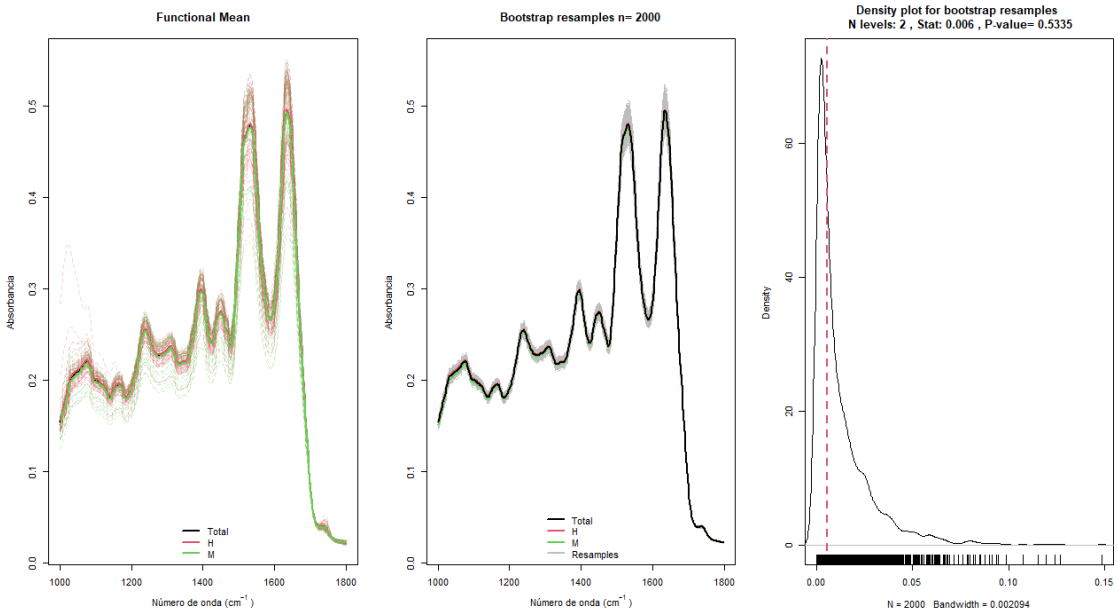
Considerando la covariable sexo, se contrastará la hipótesis nula de que la media en los hombres es igual a la media en las mujeres, esto es, $\mathcal{H}_0 : \mu_H = \mu_M$. El contraste se realiza unifactorialmente en cada región. Este, no podrá realizarse en el grupo de pacientes de cáncer de mama ya que no contiene ningún hombre.

Para el grupo de pacientes de cáncer de colon no se encuentran diferencias significativas en media, Figura 4.3, con p-valores en ambas regiones por encima de los niveles de significación usuales. Al no haber eliminado ningún atípico en este grupo se mantienen los grupos en 36 mujeres y 64 hombres, por lo que hay un ligero desbalance con las mujeres menos representadas en la submuestra.

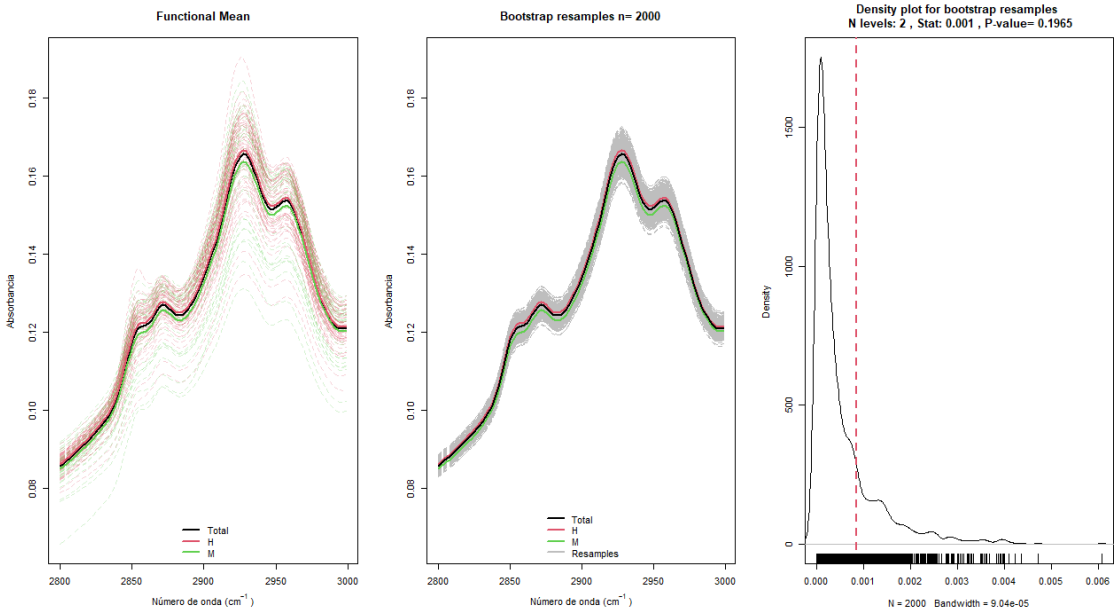
En el caso del grupo de referencia, una vez eliminada la muestra atípica quedan 57 hombres y 42 mujeres. En este caso, Figura 4.4, ambos p-valores se encuentran en los límites de significación usuales, por lo que podría intuirse una cierta diferencia entre las submuestras. En la Tabla 4.2 pueden encontrarse resumidos los p-valores obtenidos en los contrastes realizados en este apartado.

Grupo	**p-valor fingerprint**	**p-valor lípidos**
Cáncer de colon	0.53	0.20
Referencia	0.05	0.07

Tabla 4.2: Resumen de los p-valores obtenidos para los contrastes de igualdad de medias del factor sexo.

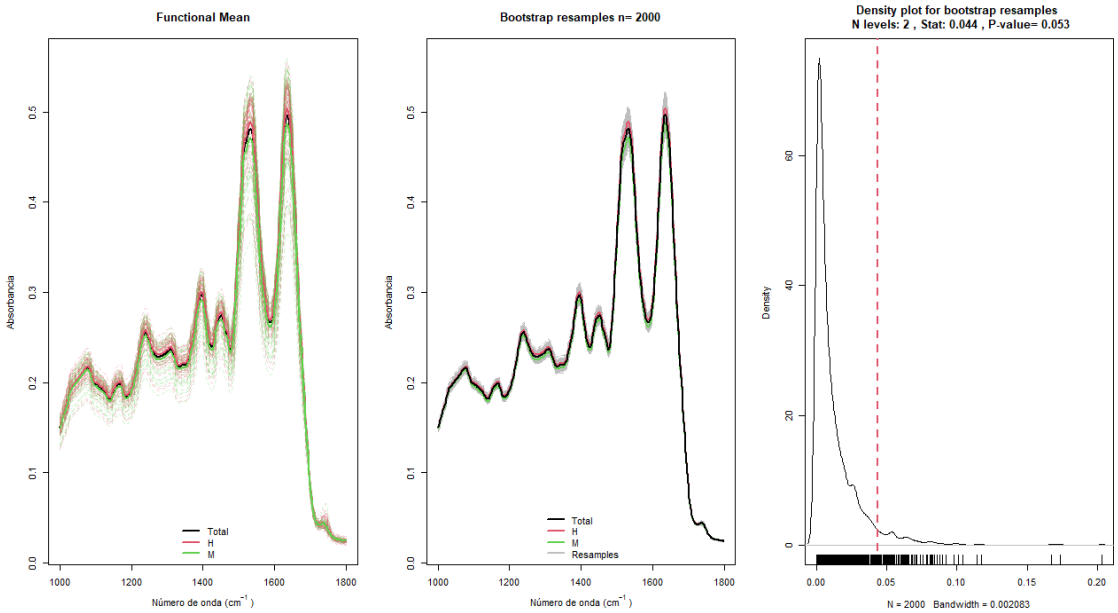


(a) *Región fingerprint*

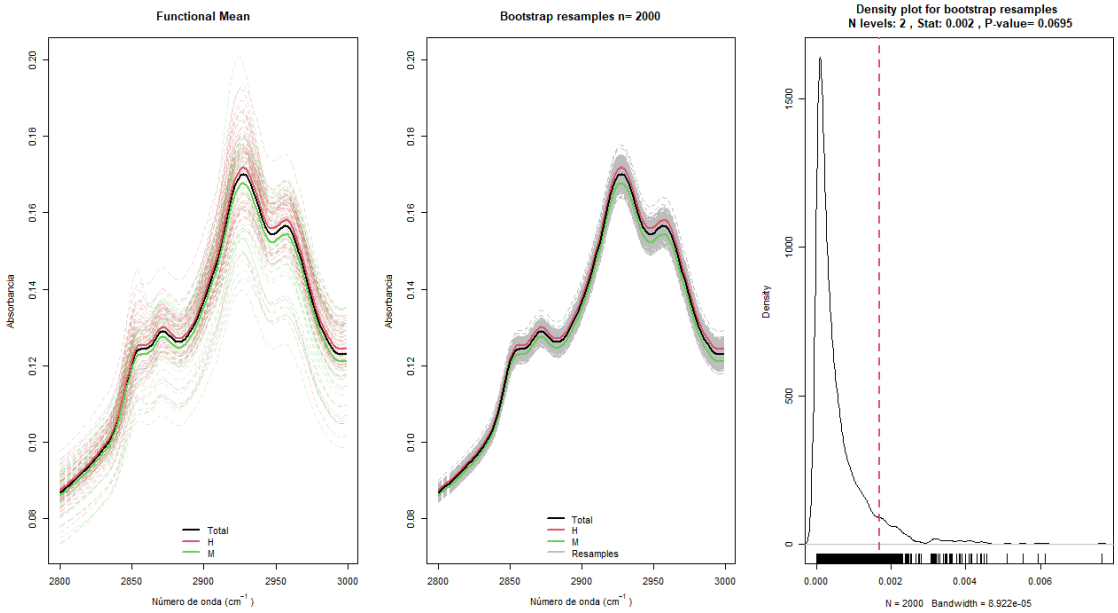


(b) *Región lípidos*

Figura 4.3: ANOVA unifactorial para el nivel sexo en el grupo de pacientes de cáncer de colon con $B = 2000$ réplicas bootstrap.



(a) *Región fingerprint*



(b) *Región lípidos*

Figura 4.4: ANOVA unifactorial para el nivel sexo en el grupo de control con $B = 2000$ réplicas bootstrap.

4.1.4. ANOVA multifactorial

El análisis de varianza multifactorial permite explorar las interacciones entre factores, revelando dependencias entre efectos. También posibilita la comparación entre múltiples grupos para ayudar a identificar diferencias significativas que podrían quedar ocultas al realizar el contraste unifactorial. Además, de esta manera se pueden aislar los efectos de interés, detectando factores de confusión y reduciendo su impacto en el análisis.

Se estudiará si las diferencias entre los espectros de hombres y mujeres varían según los grupos de pertenencia en el caso de cáncer de colon y del grupo de referencia. En las Tablas 4.3 y 4.4 se observan los p-valores obtenidos en las pruebas realizadas según 30 proyecciones aleatorias y con $B = 2000$ remuestras Bootstrap. Quedando todos los valores por encima de los niveles usuales de significación salvo por el factor grupo según el se detectan diferencias significativas según los tres ajustes del p-valor y en ambas regiones. Cabe esperar así que el sexo no sea un factor de confusión en este problema y mostrando que no tiene relación con el grupo.

p-valor	**Factor**	**Región fingerprint**	**Región lípidos**
Bonferroni	Sexo	1	1
	Grupo	0	0.06
FDR	Sexo	0.97	0.78
	Grupo	0	0.04
Bootstrap	Sexo	0.81	0.48
	Grupo	0	0.01

Tabla 4.3: Tabla resumen de los p-valores obtenidos en el estudio de los factores de sexo y grupo a la vez en el grupo de colon con el grupo de referencia utilizando 30 proyecciones aleatorias y $B = 2000$ réplicas bootstrap.

En resumen, se han detectado diferencias significativas para el factor sexo en el grupo de referencia, que es el más uniforme en cuanto a equilibrio de sexos y edad más concentrada en sus individuos. Esto se tendrá en cuenta para las tareas de clasificación retirando a los hombres del grupo de referencia en el problema de clasificación binario de cáncer de mama contra sano, evitando así la acción de variables de confusión. También se detectan diferencias significativas en media entre los grupos de referencia y los de pacientes de cáncer, lo que permite suponer que los clasificadores consigan discriminar entre los grupos de la manera en que se espera.

p-valor	**Factor**	**Región fingerprint**	**Región lípidos**
Bonferroni	Sexo	1	1
	Grupo	0	0.01
	Interacción	1	1
FDR	Sexo	0.99	0.51
	Grupo	0	0.01
	Interacción	0.43	0.70
Bootstrap	Sexo	0.56	0.54
	Grupo	0	0.01
	Interacción	0.40	0.71

Tabla 4.4: Tabla resumen de los p-valores obtenidos en el estudio de los factores de sexo, grupo y su interacción a la vez en el grupo de colon con el grupo de referencia utilizando 30 proyecciones aleatorias y $B = 2000$ réplicas bootstrap.

4.2. Modelos de clasificación: cáncer de mama vs referencia

En esta sección, se presentan los resultados obtenidos de la clasificación de pacientes con cáncer de mama contra individuos sanos. El objetivo principal es evaluar la capacidad de distintos modelos para distinguir entre las dos clases mencionadas. Como se ha mencionado en el apartado anterior, no se utilizarán los hombres de referencia, ya que podrían introducir confusión en el modelo. De esta manera, el problema queda claramente desbalanceado. Así, se crearon remuestras artificiales a partir de las 42 de referencia mediante bootstrap suavizado hasta equilibrar el problema.

Después de realizar varias pruebas con distintos modelos y sus respectivos parámetros, se utilizó un kernel lineal para el SVM, un estimador de N-W para la estimación no paramétrica, la métrica $\mathcal{L}_2$ en GKAM y profundidad modal en el clasificador basado en profundidades. La lista de todos los clasificadores entrenados es la siguiente:

- Máquinas de soporte vectorial con núcleo lineal, SVM, esta opción multivariante utilizará como variables explicativas ambas variables funcionales representadas por sus coeficientes en una base de componentes principales, la correspondiente a la región del *fingerprint* con 3 coeficientes y la de los lípidos con 2. Esta técnica es la que mejores resultados da en la bibliografía [Huber et al., 2021, Warner, 2011].

- Árbol de decisión, Tree, se entrenó utilizando la misma representación en la base de componentes principales.

- El análisis del discriminante cuadrático, QDA, es el último de los clasificadores basados en técnicas multivariantes entrenado. Este es ampliamente utilizado en el contexto de la Química y de la espectroscopía infrarroja [Huber et al., 2021].

- Estimación no paramétrica de la probabilidad mediante el estimador de Nadaraya-Watson, NP N-W, enfoque adecuado cuando la información se encuentra en las distancias, aunque es sensible a grupos desbalanceados. Se obtuvo el parámetro óptimo de la ventana h mediante una rejilla de búsqueda y se seleccionó el mejor valor mediante validación cruzada.

- Estimación no paramétrica mediante k vecinos más próximos, kNN, en donde el parámetro k de vecinos próximos se eligió también optimizando los valores de una rejilla por validación cruzada.

- Regresión binaria mediante el modelo general lineal, GLM, utilizando de nuevo la representación de los datos en la base de componentes principales en ambas variables. Este modelo es adecuado cuando las relaciones entre los grupos y las covariables son lineales.

- Regresión binaria con modelos aditivos espectrales generalizados, GSAM, considerando flexibles las contribuciones de los coeficientes de los datos. En este modelizado y en GLM la información estará recogida por al base.

- Regresión binaria con modelos aditivos kernel generalizados, GKAM, utilizando enfoques kernel con la métrica de las covariables como contribuciones flexibles de los coeficientes de la base. En este modelo la información está recogida en las métricas de las covariables.

- DD-plot con modelo aditivo generalizado, DD-GAM, basándose en el concepto de profundidades se representan las observaciones según su profundidad modal respecto a los grupos y

se modelan dichas profundidades mediante una regresión flexible para realizar clasificaciones [Cuevas et al., 2007].

Se entrenan los clasificadores presentados utilizando un 75 % de la muestra como entrenamiento y el 25 % restante como *test* para evaluar el rendimiento de los modelos. La muestra de entrenamiento se elige de manera aleatoria y se repite este proceso de entrenamiento y evaluación $N = 100$ veces para conseguir resultados más generales. Los valores de las métricas de evaluación utilizadas: Acc, TPR, TNR, PPV y F1 se encuentran resumidos en las Tablas 4.5, 4.6, 4.7, 4.8 y 4.9 respectivamente. Cabe mencionar que todos los clasificadores optimizan la *accuracy*, pero al incluir las demás métricas se logra una una visión más completa del comportamiento de los modelos.

Técnica	$\overline{\textbf{Acc}}$	**mín(Acc)**	**máx(Acc)**	$\boldsymbol{\sigma}$
SVM	0.87	0.76	0.96	0.04
Tree	0.79	0.69	0.94	0.05
QDA	0.92	0.86	1.00	0.03
NP N-W	0.86	0.73	0.94	0.05
kNN	0.93	0.82	1.00	0.04
GLM	0.88	0.78	0.96	0.04
GSAM	0.85	0.73	0.96	0.05
GKAM	0.91	0.82	0.98	0.04
DD-GAM	0.83	0.67	0.94	0.06

Tabla 4.5: Valores de Acc media, mínima, máxima y desviación estándar al realizar $N = 100$ entrenamientos y evaluaciones de los clasificadores probados para el problema de clasificación cáncer de mama contra grupo de referencia. En amarillo aparece destacado el mayor valor medio de Acc, alcanzado por el clasificador kNN.

En los valores de sensibilidad destaca los valores de desviación típica del árbol de decisión y del clasificador basado en profundidades, que también cuentan con mayor rango de valores, lo que indica que no son estables. Esto es problemático puesto que muestra dependencia en los datos de entrenamiento, restándole fiabilidad y capacidad de generalización al método.

Técnica	$\overline{\textbf{TPR}}$	**mín(TPR)**	**máx(TPR)**	$\boldsymbol{\sigma}$
SVM	0.89	0.71	1.00	0.06
Tree	0.81	0.47	1.00	0.11
QDA	0.93	0.81	1.00	0.05
NP N-W	0.91	0.74	1.00	0.05
kNN	0.91	0.70	1.00	0.06
GLM	0.92	0.81	1.00	0.04
GSAM	0.88	0.64	1.00	0.07
GKAM	0.90	0.74	1.00	0.06
DD-GAM	0.85	0.62	1.00	0.09

Tabla 4.6: Valores de TPR medio, mínimo, máximo y desviación estándar al realizar $N = 100$ entrenamientos y evaluaciones de los clasificadores probados para el problema de clasificación cáncer de mama contra grupo de referencia. En amarillo aparece sombreada la celda de mayor valor medio de TPR, que se corresponde con el clasificador multivariante QDA.

En cuanto a los valores de especificidad, Tabla 4.7, destaca de nuevo la inestabilidad del árbol de decisión y de DD-GAM, esto pudo camuflarse para los valores de *accuracy* compensándose con la sensibilidad, por lo que puede ser que en ciertas ocasiones esté sobreajustando a una clase y en otras a la otra.

Técnica	$\overline{\textbf{TNR}}$	**mín(TNR)**	**máx(TNR)**	$\boldsymbol{\sigma}$
SVM	0.84	0.64	1.00	0.08
Tree	0.77	0.52	0.97	0.10
QDA	0.92	0.80	1.00	0.05
NP N-W	0.81	0.52	1.00	0.09

Técnica	$\overline{\textbf{TNR}}$	**mín(TNR)**	**máx(TNR)**	$\boldsymbol{\sigma}$
kNN	0.94	0.79	1.00	0.05
GLM	0.85	0.64	0.96	0.07
GSAM	0.83	0.58	1.00	0.09
GKAM	0.92	0.73	1.00	0.06
DD-GAM	0.81	0.52	1.00	0.09

Tabla 4.7: Valores de TNR medio, mínimo, máximo y desviación estándar al realizar $N = 100$ entrenamientos y evaluaciones de los clasificadores probados para el problema de clasificación cáncer de mama contra grupo de referencia. En amarillo aparece sombreada la celda de mayor valor medio de TNR, que se corresponde con kNN.

Técnica	$\overline{\textbf{PPV}}$	**mín(PPV)**	**máx(PPV)**	$\boldsymbol{\sigma}$
SVM	0.85	0.67	1.00	0.07
Tree	0.78	0.62	0.96	0.08
QDA	0.92	0.77	1.00	0.05
NP N-W	0.83	0.63	1.00	0.08
kNN	0.94	0.76	1.00	0.05
GLM	0.86	0.69	0.97	0.06
GSAM	0.84	0.65	1.00	0.08
GKAM	0.92	0.75	1.00	0.06
DD-GAM	0.81	0.63	1.00	0.08

Tabla 4.8: Valores de precisión de los clasificadores probados para el problema de clasificación cáncer de mama contra grupo de referencia al realizar $N = 100$ entrenamientos y evaluaciones. En amarillo aparece sombreada la celda de mayor valor medio de precisión, que se corresponde con el clasificador kNN.

Técnica	$\overline{\textbf{F1}}$	**mín(F1)**	**máx(F1)**	$\boldsymbol{\sigma}$
SVM	0.87	0.74	0.95	0.05
Tree	0.79	0.62	0.94	0.06
QDA	0.92	0.85	1.00	0.03
NP N-W	0.87	0.73	0.95	0.05
kNN	0.92	0.80	1.00	0.04
GLM	0.89	0.76	0.97	0.04
GSAM	0.85	0.68	0.97	0.05
GKAM	0.91	0.78	0.98	0.04
DD-GAM	0.83	0.67	0.95	0.06

Tabla 4.9: Valores de F1 de los clasificadores probados para el problema de clasificación cáncer de mama contra grupo de referencia al realizar $N = 100$ entrenamientos y evaluaciones. En amarillo aparece sombreada la celda de mayor valor medio de la métrica F1, que se corresponde con el clasificador kNN.

En general se obtienen resultados alentadores, siendo el clasificador basado en k vecinos más próximos, KNN, el de mejores resultados consistentemente en casi todas las métricas. Este discriminador sólo se ve superado en la sensibilidad por el análisis del discriminante cuadrático. Al ser la sensibilidad la tasa de verdaderos positivos, esto puede indicar que kNN es más conservador a la hora de clasificar una nueva observación como positiva.

Los clasificadores que obtuvieron los menores valores medios de las métricas de evaluación fueron el árbol de decisión (Tree) y el basado en profundidades (DD-GAM). Teniendo, además, las mayores desviaciones típicas, indicando que son inestables y que dependen de la muestra de entrenamiento, algo no deseable en este contexto en el que se busca generalidad y poder de escalabilidad. De los valores observados en las tablas se deduce que si se comportan bien en sensibilidad se comportan mal en

especificidad, por lo tanto estarán siempre clasificando mayoritariamente según un único grupo. Para el DD-plot, esto puede deberse a que se advierte dispersión en las nubes de puntos, reflejada de la dispersión en los datos originales y a que no aparecen gráficos con forma de L en los que las muestras estarían separadas, Figura 4.5.

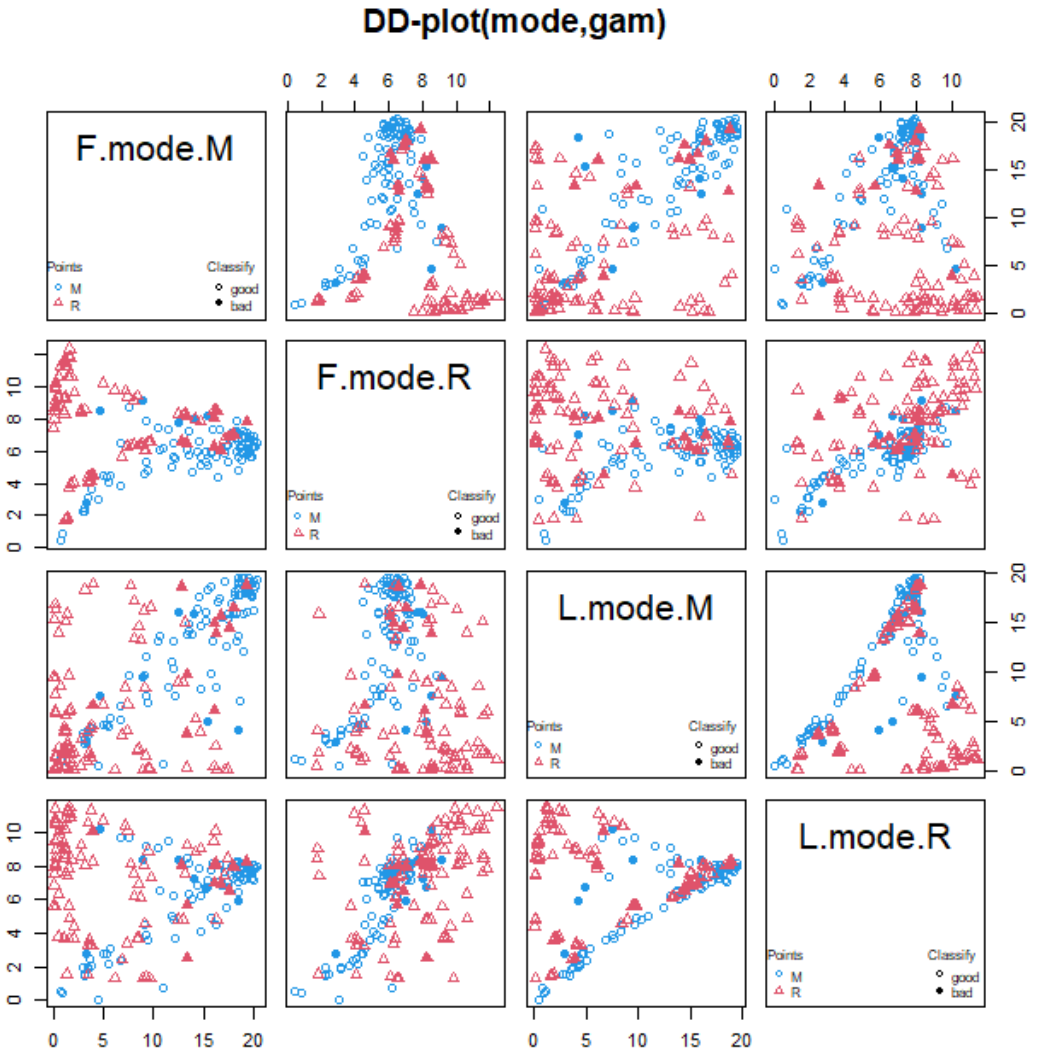


Figura 4.5: DD-plots de los grupos de cáncer de mama y de referencia considerando la profundidad modal.

4.3. Modelos de clasificación: cáncer de colon vs referencia

En esta sección, se presentan los resultados obtenidos para el problema de clasificación binario entre cáncer de colon y grupo de referencia. El procedimiento es el mismo que en el caso anterior del

cáncer de mama, entrenando iterativamente los mismos clasificadores presentados en la sección anterior y evaluándolos según distintas métricas para poder compararlos. Los resultados se encuentran en las Tablas 4.10, 4.11, 4.12, 4.11 y 4.14. Sin embargo, en este caso se utilizaron directamente las muestras originales, sin ser necesario crear remuestras.

En cuanto a los valores de *accuracy* obtenidos, en general, son algo más bajos en media que en la sección anterior. Destaca que para el modelo basado en estimación de la densidad NP N-W y el basado en DD-plot se alcanzaron valores mínimos por debajo de 0.5, por lo que serían peores que un estimador *naive* que clasificara todas las observaciones según un único grupo, lo que no es deseable.

Técnica	$\overline{\textbf{Acc}}$	**mín(Acc)**	**máx(Acc)**	$\boldsymbol{\sigma}$
SVM	0.84	0.72	0.94	0.05
Tree	0.77	0.66	0.92	0.05
QDA	0.90	0.72	0.98	0.05
NP N-W	0.69	0.46	0.84	0.08
kNN	0.77	0.66	0.90	0.06
GLM	0.77	0.64	0.86	0.05
GSAM	0.75	0.62	0.86	0.05
GKAM	0.67	0.50	0.80	0.07
DD-GAM	0.68	0.46	0.84	0.07

Tabla 4.10: Valores de Acc media, mínima, máxima y desviación estándar al realizar $N = 100$ entrenamientos y evaluaciones de los clasificadores probados para el problema de clasificación cáncer de colon contra referencia. En amarillo aparece destacado el mayor valor medio de la Acc, alcanzado por el clasificador QDA.

Técnica	$\overline{\textbf{TPR}}$	**mín(TPR)**	**máx(TPR)**	$\boldsymbol{\sigma}$
SVM	0.86	0.65	1.00	0.08
Tree	0.77	0.48	1.00	0.11

Técnica	$\overline{\textbf{TPR}}$	**mín(TPR)**	**máx(TPR)**	$\boldsymbol{\sigma}$
QDA	0.86	0.56	1.00	0.08
NP N-W	0.77	0.46	1.00	0.10
kNN	0.78	0.52	0.96	0.09
GLM	0.77	0.50	0.96	0.09
GSAM	0.75	0.50	0.92	0.09
GKAM	0.74	0.39	1.00	0.12
DD-GAM	0.70	0.42	0.92	0.10

Tabla 4.11: Valores de TPR medio, mínimo, máximo y desviación estándar al realizar $N = 100$ entrenamientos y evaluaciones de los clasificadores probados para el problema de clasificación cáncer de colon contra grupo de referencia. En amarillo aparece sombreada la celda de mayor valor medio de TPR, que se corresponde con el clasificador multivariante SVM.

La inestabilidad en los resultados de los modelos se observa de forma transversal en todas las métricas, siendo el QDA el modelo más robusto en el sentido de que tiene los menores valores de desviación estándar en sus resultados.

Técnica	$\overline{\textbf{TNR}}$	**mín(TNR)**	**máx(TNR)**	$\boldsymbol{\sigma}$
SVM	0.81	0.59	1.00	0.08
Tree	0.76	0.52	0.96	0.10
QDA	0.94	0.72	1.00	0.05
NP N-W	0.61	0.30	0.92	0.15
kNN	0.76	0.46	0.96	0.10
GLM	0.77	0.50	1.00	0.09
GSAM	0.75	0.48	0.96	0.09

Técnica	$\overline{\textbf{TNR}}$	**mín(TNR)**	**máx(TNR)**	$\boldsymbol{\sigma}$
GKAM	0.62	0.27	0.92	0.15
DD-GAM	0.66	0.42	0.88	0.10

Tabla 4.12: Valores de TNR medio, mínimo, máximo y desviación estándar al realizar $N = 100$ entrenamientos y evaluaciones de los clasificadores probados para el problema de clasificación cáncer de colon contra grupo de referencia. En amarillo aparece sombreada la celda de mayor valor medio de TNR, que se corresponde con QDA.

Técnica	$\overline{\textbf{PPV}}$	**mín(PPV)**	**máx(PPV)**	$\boldsymbol{\sigma}$
SVM	0.83	0.65	1.00	0.07
Tree	0.77	0.56	0.96	0.09
QDA	0.93	0.71	1.00	0.06
NP N-W	0.67	0.37	0.92	0.12
kNN	0.77	0.57	0.96	0.09
GLM	0.77	0.60	1.00	0.07
GSAM	0.76	0.53	0.94	0.08
GKAM	0.67	0.40	0.89	0.12
DD-GAM	0.66	0.44	0.84	0.08

Tabla 4.13: Valores de precisión medios, mínimos, máximos y deviaciones estándar al realizar $N = 100$ entrenamientos y evaluaciones de los clasificadores probados para el problema de clasificación cáncer de colon contra grupo de referencia. En amarillo aparece sombreada la celda de mayor valor medio de precisión, que se corresponde con el clasificador QDA.

Técnica	$\overline{\textbf{F1}}$	**mín(F1)**	**máx(F1)**	$\boldsymbol{\sigma}$
SVM	0.84	0.68	0.94	0.05
Tree	0.76	0.60	0.90	0.06
QDA	0.89	0.72	0.98	0.05
NP N-W	0.71	0.49	0.84	0.07
kNN	0.77	0.62	0.89	0.06
GLM	0.76	0.60	0.88	0.06
GSAM	0.75	0.56	0.88	0.06
GKAM	0.69	0.54	0.82	0.07
DD-GAM	0.67	0.45	0.83	0.08

Tabla 4.14: Valores de F1 de los clasificadores probados para el problema de clasificación cáncer de colon contra grupo de referencia al realizar $N = 100$ entrenamientos y evaluaciones. En amarillo aparece sombreada la celda de mayor valor medio de la métrica F1, que se corresponde con el clasificador QDA.

Los resultados obtenidos representan una base sólida que respalda el potencial uso de la espectroscopía infrarroja para la detección de cáncer. De nuevo, el clasificador con mejor rendimiento consistentemente en la mayoría de las métricas fue el QDA, salvo para la sensibilidad, que coincide con el SVM. Así, los clasificadores con mejores resultados fueron multivariantes, esto puede indicar que la base de componentes principales resume bien la información de la muestra y sirve para discriminar los grupos. Así mismo, se observa una inestabilidad en el árbol de decisión (Tree), los clasificadores basados en kernel y en regresión binaria, teniendo valores mínimos por debajo de 0.50 en algunos casos en los que serían peores que un clasificador *naive*, por lo que no tienen un comportamiento adecuado. Esto puede deberse a que no se estén optimizando bien sus parámetros. En cuanto a los clasificadores basados en profundidades, sus malos resultados pueden ser debidos al gran solape entre los dos grupos que se observa en la tendencia diagonal de los gráficos de dispersión en la Figura 4.6.

Cabe mencionar que, pese a obtener buenos resultados en ambos casos, parece que se fueron mejores los del caso de cáncer de mama que los del cáncer de colon. Esto podría derivarse a que el cáncer de mama tenga una mayor respuesta en sangre que el cáncer de colon, o a que, pese a no esperarlo, mezclar los sexos esté generando confusión en el problema.

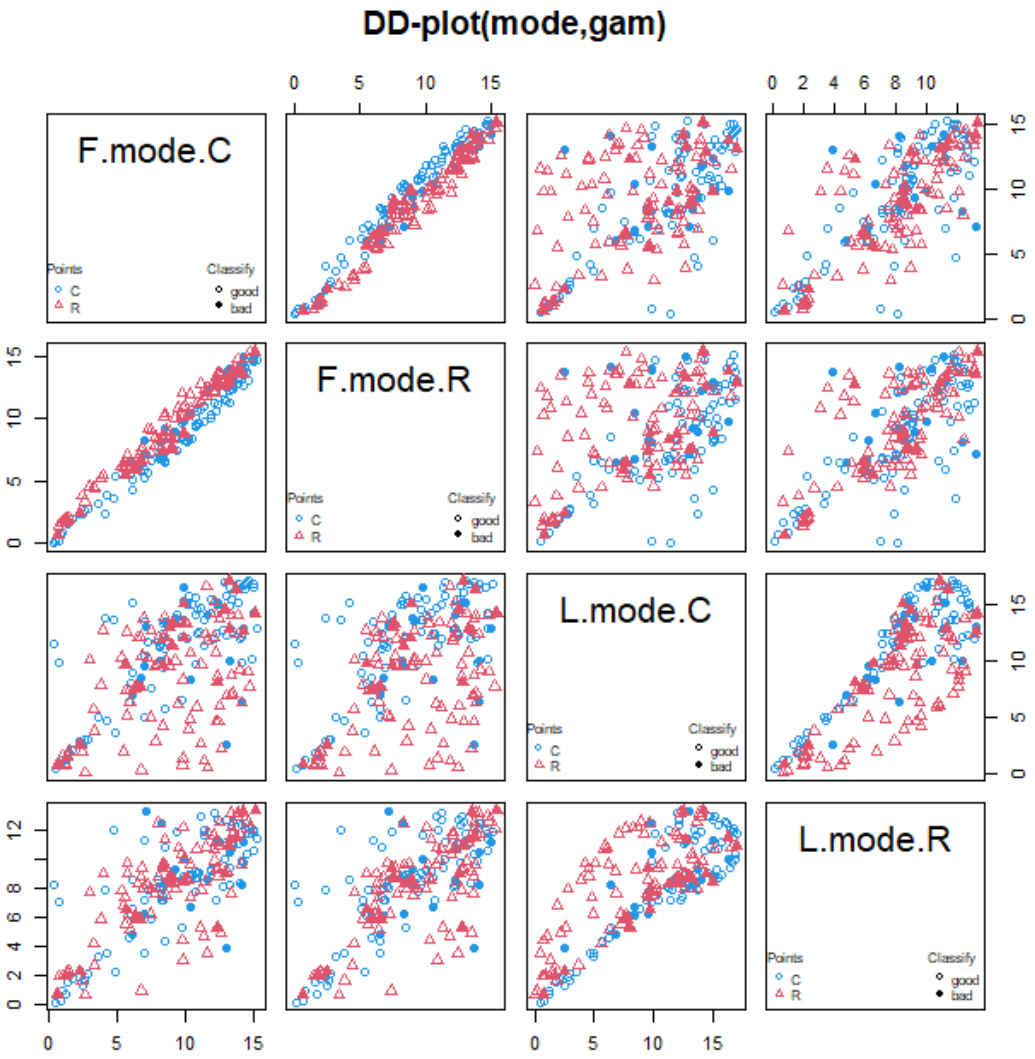


Figura 4.6: DD-plots de los grupos de cáncer de colon y de referencia considerando la profundidad modal.

Conclusiones

El principal objetivo de este trabajo era el de verificar la capacidad de la espectroscopía infrarroja como método de detección de cáncer en muestras de suero sanguíneo. Este estudio ha sido motivado por la necesidad de desarrollar técnicas de diagnóstico más rápidas, precisas y no invasivas que puedan complementar o mejorar los métodos actuales, contribuyendo así a una detección temprana, mejor pronóstico y mayor tasa de participación en los cribados.

Para poder realizar este estudio, en primer lugar, se ha realizado una aproximación formal a las nociones básicas de teoría de análisis de datos funcionales necesarias para comprender el contexto y la metodología empleada. Esto incluye conceptos desde la representación de datos funcionales, métricas y espacios funcionales, medidas de localización, pasando por técnicas de visualización como el análisis de componentes principales funcional, hasta la presentación de modelos como ANOVA unifactorial y multifactorial y distintos tipos de modelos de clasificación.

Pasando al análisis de la base de datos, se contaba con una base de espectros infrarrojos de pacientes con cáncer de mama, colon e individuos sanos de referencia. Sobre esta base de datos se realizó un análisis exploratorio, con el objetivo de comprender la estructura subyacente de los datos y su variabilidad. Se representaron medidas de localización central para cada grupo, se detectaron y eliminaron datos atípicos y se realizó un análisis de las componentes principales funcional.

A continuación, antes de abordar el problema de clasificación binaria entre el grupo de pacientes con cáncer de mama y el grupo sano de referencia, así como su análogo con pacientes de cáncer de colon, se llevó a cabo un análisis de la varianza para detectar posibles factores adicionales que pudieran influir en el problema, además del factor grupo. Este análisis reveló que el sexo podría ser un factor relevante, por lo que se tuvo en cuenta a la hora de seleccionar el grupo de referencia para la clasificación. También se detecto un sesgo en la edad en cada grupo, por lo que podría ser una variable de confusión. Se aplicaron modelos de clasificación multivariantes, modelos basados en estimación kernel, modelos basados en regresión binaria y clasificadores basados en profundidades. Se evaluaron según las métricas *accuracy*, sensibilidad, especificidad, precisión y F1, concluyendo que el clasificador con mejor rendimiento en el problema de discriminar cáncer de mama del grupo de referencia fue el kNN, obteniendo el mejor resultado en la mayoría de las métricas **??**.

Por otra parte, en el caso de clasificar cáncer de colon y grupo de referencia los mejores resultados se obtuvieron en la mayoría de las métricas con el análisis del discriminante cuadrático, QDA, Tabla 4.16.

Los resultados obtenidos en este trabajo son alentadores y demuestran el potencial de la espectroscopía infrarroja como herramienta para la detección de cáncer. Sin embargo, cabría mencionar las siguientes limitaciones dentro del estudio

$\overline{\textbf{Acc}}$	$\overline{\textbf{TPR}}$	$\overline{\textbf{TNR}}$	$\overline{\textbf{PPV}}$	$\overline{\textbf{F1}}$
0.92	0.93	0.94	0.94	0.92

Tabla 4.15: Valores medios de las métricas de evaluación de kNN en el problema de clasificación de cáncer de mama contra grupo de referencia.

$\overline{\textbf{Acc}}$	$\overline{\textbf{TPR}}$	$\overline{\textbf{TNR}}$	$\overline{\textbf{PPV}}$	$\overline{\textbf{F1}}$
0.90	0.86	0.94	0.93	0.89

Tabla 4.16: Valores medios de las métricas de evaluación de QDA en el problema de clasificación de cáncer de colon contra grupo de referencia.

- El cáncer es una enfermedad muy compleja, que produce distintas respuestas según su grado, estadio y subtipos. Se podría estudiar si estas variables representan factores en el problema. En este trabajo, debido a que algunos diagnósticos estaban incompletos, se estudiaron agrupados en mama y colon.
- Es posible que la eliminación de posibles atípicos haya sido demasiado estricta, lo que podría haber llevado a la exclusión de datos no atípicos. Un enfoque más conservador para la eliminación podría haber sido una mejor opción para conservar información.
- La base de datos actual es de tamaño limitado, lo que restringe el estudio de la escalabilidad y generalización de los resultados presentados en este estudio.
- No se han explorado las derivadas espectrales en este estudio, lo que limita la capacidad de estudiar proporciones de concentraciones y de eliminar posibles sesgos debidos a la escala, así como de comprobar la robustez de los resultados presentados.

Apéndice A

Código del Análisis Exploratorio

```
# ANALISIS EXPLORATORIO
# Se cargan los datos, primero con la zona del fingerprint, F, (1000-1800 cm^{-1})
mdata = rbind(mamaF,colonF, controlF)
# Defino la variable grupo en la que M es cancer de mama, C es de colon y R es de referencia
grup = as.factor(c(rep("M",100),rep("C",100),rep("R",100)))
# Se usarán los colores azul:dodgerblue4 , amarillo:yellow1 y rojo: red4.

# Se crea el objeto fdata
fingerprint = fdata(as.matrix(mdata), argvals=as.matrix(lambda1),rangeval=c(lambda1[1,1],
    lambda1[415,length(lambda1)])) # mdata menos primera columna
fingerprint$names$xlab = expression(paste("Número de onda (",cm^{-1}, " )"))
fingerprint$names$main = ""
fingerprint$names$ylab = "Absorbancia"
plot(fingerprint[grup == "M"],xlab=expression(paste("Número de onda (",cm^{-1}, " )")),ylab=
    "Absorbancia",col="red4",ylim=c(0,0.55))
lines(fingerprint[grup == "C"],col="yellow3")
lines(fingerprint[grup == "R"],col="dodgerblue4")
legend("topleft",c("Cáncer de mama","Cáncer de colon", "Grupo de referencia"),col=c("red4","
    yellow3","dodgerblue4"),lwd=1)

# Se crea ahora la variable lípidos (con el identificativo L)
mdata=rbind(mamaL,colonL, controlL)
lipidos=fdata(as.matrix(mdata), argvals=as.matrix(lambda2),rangeval=c(lambda2[1,1],lambda2
    [104,length(lambda2)]))
lipidos$names$xlab=expression(paste("Número de onda (",cm^{-1}, " )"))
lipidos$names$main=""
lipidos$names$ylab="Absorbancia"
plot(lipidos[grup=="M"],xlab=expression(paste("Número de onda (",cm^{-1}, " )")),ylab="
    Absorbancia",col="red4",ylim=c(0,0.55),xlim=c(2200,3000))
lines(lipidos[grup=="C"],col="yellow3")
lines(lipidos[grup=="R"],col="dodgerblue4")
legend("topright",c("Cáncer de mama","Cáncer de colon", "Grupo de referencia"),col=c("red4",
    "yellow3","dodgerblue4"),lwd=1)

# Medidas de localización central
# Se calcula la media en cada grupo y en cada variable (fingerprint y lípidos)
mediaM_F = fdata.bootstrap(fingerprint[grup =='M'], statistic = func.mean, nb = 500,draw=
    TRUE)
mediaC_F = fdata.bootstrap(fingerprint[grup == "C"], statistic = func.mean, nb = 500,draw=
    TRUE)
mediaR_f = fdata.bootstrap(fingerprint[grup == "R"], statistic = func.mean, nb = 500,draw=
    TRUE)

mediaM_L = fdata.bootstrap(lipidos[grup == 'M' ], statistic = func.mean, nb = 500,draw=TRUE)
```

```
36 mediaC_L = fdata.bootstrap(lipidos[grup == 'C' ], statistic = func.mean, nb = 500,draw=TRUE)
37 mediaR_L = fdata.bootstrap(lipidos[grup == 'R' ], statistic = func.mean, nb = 500,draw=TRUE)
38
39 # Se calcula la mediana con distintas profundidades
40 mc = c("red4", "yellow3", "dodgerblue4") # Se crea un vector con el código de color
41 # Utilizando la profundidad RP en ambas variables y 15 proyecciones
42 mamaRP_F = depth.RP(fingerprint[grup == 'M'],nproj=15)
43 colonR_PF = depth.RP(fingerprint[grup == 'C'],nproj=15)
44 refRP_F = depth.RP(fingerprint[grup == 'R'],nproj=15)
45
46 mamaRP_L = depth.RP(lipidos[grup == 'M'],nproj=15)
47 colonRP_L = depth.RP(lipidos[grup == 'C'],nproj=15)
48 refRP_L = depth.RP(lipidos[grup == 'R'],nproj=15)
49
50 # Utilizando la profundidad modal en ambas variables
51 mamamode_F = depth.mode(fingerprint[grup == 'M'])
52 colonmode_F = depth.mode(fingerprint[grup == 'C'])
53 refmode_F = depth.mode(fingerprint[grup == 'R'])
54
55 mamamode_L = depth.mode(lipidos[grup == 'M'])
56 colonmode_L = depth.mode(lipidos[grup == 'C'])
57 refmode_L = depth.mode(lipidos[grup == 'R'])
58
59 # Atípicos
60 # Se estudiarán en cada grupo, primero con la metodología bootstrap ponderado
61 out.FMw.ma_F = outliers.depth.pond(fingerprint[grup == 'M'], nb=1000, smo=0.1, dfunc=depth.
       FM)
62 out.FMw.ma_L = outliers.depth.pond(lipidos[grup == 'M'], nb=1000, smo=0.1, dfunc=depth.FM)
63
64 out.FMw.col_F = outliers.depth.pond(fingerprint[grup == 'C'], nb=1000, smo=0.1, dfunc=depth.
       FM)
65 out.FMw.col_L = outliers.depth.pond(lipidos[grup == 'C'], nb=1000, smo=0.1, dfunc=depth.FM)
66
67 out.FMw.ref_F = outliers.depth.pond(fingerprint[grup == 'R'], nb=1000, smo=0.1, dfunc=depth.
       FM)
68 out.FMw.ref_L = outliers.depth.pond(lipidos[grup == 'R'], nb=1000, smo=0.1, dfunc=depth.FM)
69
70 # Se repite con el procedimiento bootstrap recortado y la profundidad modal
71 out.modet.ma_F = outliers.depth.pond(fingerprint[grup == 'M'], nb=1000, smo=0.1, dfunc=depth
       .mode)
72 out.modet.ma_L = outliers.depth.pond(lipidos[grup == 'M'], nb=1000, smo=0.1, dfunc=depth.
       mode)
73
74 out.modet.col_F = outliers.depth.pond(fingerprint[grup == 'C'], nb=1000, smo=0.1, dfunc=
       depth.mode)
75 out.modet.col_L = outliers.depth.pond(lipidos[grup == 'C'], nb=1000, smo=0.1, dfunc=depth.
       mode)
76
77 out.modet.ref_F = outliers.depth.pond(fingerprint[grup == 'R'], nb=1000, smo=0.1, dfunc=
       depth.mode)
78 out.modet.ref_L = outliers.depth.pond(lipidos[grup == 'R'], nb=1000, smo=0.1, dfunc=depth.
       mode)
```

Apéndice B

Código ANOVA y clasificación

B.1. ANOVA

```
# ANOVA
# Se realiza el ANOVA unifactorial de grupo
# Se crean las remuestras de las mujeres del grupo de referencia, que servirán también luego en la clasificación
control_mamaF = fingerprint[grup =='R']
control_mamaL = lipidos[grup == 'R']
ind_controlm = grupos=='M' # grupos es una variable con el sexo de los individuos de control
control_mamaF = control_mamaF[ind_controlm]
control_mamaL = control_mamaL[ind_controlm]

B <- 56  # tamaño de la remuestra (hasta equilibrar las clases de control y cáncer de mama)
indices <- 1:42 # índices una vez seleccionadas las mujeres del grupo de control
resample_ind <- sample(indices, size = B, replace = TRUE) # remuestreamos con reemplazamiento

# Se crean los vectores media para el ruido en ambas variables
mediaF <- rep(0, dim(fingerprint)[2])
mediaL <- rep(0,dim(lipidos)[2])
h <- 0.05 # Parámetro de suavizado
# Matrices de varianza en cada variable escaladas por el parámetro h
varF <- var(control_mamaF[['data']]) * h
varL <- var(control_mamaL[['data']]) * h

# Se genera ruido según un proceso Gaussiano de media cero y varianza proporcional a la de los datos
ruidoF = rproc2fdata(n = 56, lambda1$V1, mu = mediaF, sigma = h*varF)
ruidoL = rproc2fdata(n = 56, lambda2$V1, mu = mediaL, sigma = h*varL)

resample_F <- control_mamaF[resample_ind] + ruidoF
resample_L <- control_mamaL[resample_ind] + ruidoL

# Se concatenan los datos originales con los remuestreados
control_mujF <- fdata(as.matrix(rbind(control_mamaF$data,resample_F$data)), argvals=as.matrix(lambda1),rangeval=c(lambda1[1,1],lambda1[415,length(lambda1)]))
control_mujL <- fdata(as.matrix(rbind(control_mamaL$data,resample_L$data)), argvals=as.matrix(lambda2),rangeval=c(lambda2[1,1],lambda2[104,length(lambda2)]))

# Ahora ya se puede hacer el ANOVA
ind_macol <- grup =='M' | grup == 'R'
datafanovaF <- c(fingerprint[grup=='M'],control_mujF)
datafanovaL <- c(lipidos[grup=='M'],control_mujL)
```

```
37 res.ma.g_F <- fanova.onefactor(datafanovaF, grup[ind_macol], nboot = 2000, plot=TRUE)
38 res.ma.g_L <- fanova.onefactor(datafanovaL, grup[ind_macol], nboot = 2000, plot=TRUE)
39
40 # Se estudia el factor grupo para el cáncer de colon
41 ind_colcon = grup=='C' | grup=='R'
42 res.col.g_F = fanova.onefactor(fingerprint[ind_colcon], grup[ind_colon], nboot=2000, plot=
       TRUE)
43 res.col.g_L = fanova.onefactor(lipidos[ind_colcon], grup[ind_colon], nboot=2000, plot=TRUE)
44
45
46 # Se realiza el ANOVA del factor sexo
47 # Dentro del grupo de pacientes de cáncer de colon y en ambas variables
48 colon_sexo=read.csv2("genero_colon.csv",sep=";",dec=".",header=TRUE)
49 grupos_colon=as.factor(colon_sexo$Sexo)
50 summary(grupos_colon)
51 res.col.s_F = fanova.onefactor(fingerprint[grup == 'C'], grupos_colon, nboot=2000, plot=TRUE
       )
52 res.col.s_L = fanova.onefactor(lipidos[grup == 'C'], grupos_colon, nboot=2000, plot=TRUE)
53
54 # Dentro del grupo de control en la varaible de fingerprint y de lípidos
55 grupos = as.factor(grupos) # variable factor que contiene los sexos de los individuos de
       control
56 summary(grupos)
57 res.ref.s_F = fanova.onefactor(fingerprint[grup == 'R'], grupos, nboot=2000, plot=TRUE)
58 res.ref.s_F = fanova.onefactor(lipidos[grup == 'R'], grupos, nboot=2000, plot=TRUE)
59
60
61 # ANOVA bifactorial
62 # Para hacer el FANOVA con dos factores (factor grupo y factor sexo) se usa la función
       fanova.RPm
63 # Primero en el grupo de cáncer de colon
64 sexo <- c(grupos,grupos_colon)
65 grupp <- as.factor(c(rep('C',100),rep('R',99)))
66 m0 <- data.frame(sexo,grupp)
67 datafanovaF = fingerprint[ind_colcon]
68 datafanovaL = lipidos[ind_colcon]
69 # Sin interacción
70 res.multi_F <- fanova.RPm(datafanovaF, ~sexo+grupp, m0, nboot = 2000, hetero = FALSE)
71 res.multi_L <- fanova.RPm(datafanovaL, ~sexo+grupp, m0, nboot = 2000, hetero = FALSE)
72
73 # Con interacción
74 res.multi.int_F <- fanova.RPm(datafanovaF, ~sexo+grupp+sexo:grupp, m0, nboot = 2000, hetero
       = FALSE)
75 res.multi.int_L <- fanova.RPm(datafanovaL, ~sexo+grupp+sexo:grupp, m0, nboot = 2000, hetero
       = FALSE)
```

B.2. Clasificación

```
# CLASIFICACIÓN
# Cáncer de mama vs grupo de control, se utilizarán las remuestras anteriores
# Inicializamos valores para el bucle
clas_mamaF <- c(fingerprint[grup=='M'],control_mujF)
clas_mamaL <- c(lipidos[grup=='M'],control_mujL)
g_mama <- as.factor(c(rep('M',98),rep('R',98)))
n_mama = dim(clas_mamaF)[1]

N=100 # número de iteraciones
# Se almacenarán los valores de verdaderos positivos (tp), falsos positivos (fp)
# verdaderos negativos (tn) y falsos negativos (fn) en matrices por columnas a partir
# de los cuales se calcularán las distintas métricas de evaluación
mr.svmlin <- matrix(NA, nrow = N, ncol = 4)
mr.rpart <- matrix(NA, nrow = N, ncol = 4)
mr.qda <- matrix(NA, nrow = N, ncol = 4)
mr.np <- matrix(NA, nrow = N, ncol = 4)
mr.knn <- matrix(NA, nrow = N, ncol = 4)
mr.glm <- matrix(NA, nrow = N, ncol = 4)
mr.gsam <- matrix(NA, nrow = N, ncol = 4)
mr.gkam <- matrix(NA, nrow = N, ncol = 4)
mr.dd <- matrix(NA, nrow = N, ncol = 4)


# Bucle de entrenamiento
for (i in 1:N){
  itrain = sample(n_mama,round(n_mama*0.75),replace=FALSE)
  lista1 = list(df=data.frame(g_mama=g_mama[itrain]),f=clas_mamaF[itrain],l=clas_mamaL[
      itrain])
  b.x1 = list(f=create.pc.basis(clas_mamaF[itrain],1:3),l=create.pc.basis(clas_mamaL[itrain
      ],1:2))
  test = list(df=data.frame(g_mama = g_mama[-itrain]),f=clas_mamaF[-itrain],l=clas_mamaL[-
      itrain])
  b.xt = list(f=create.pc.basis(clas_mamaF[-itrain],1:3),l=create.pc.basis(clas_mamaL[-
      itrain],1:2))

  # Clasificadores multivariantes
  res.svmlin = classif.svm(g_mama~f+l,data=lista1, kernel ='linear')
  res.rpart = classif.rpart(g_mama~f+l,data=lista1)
  res.qda = classif.qda(g_mama~f+l,data=lista1)

  # Clasificación tipo kernel
  res.np = classif.np(g_mama[itrain],clas_mamaF[itrain], h = seq(0.1, 0.7, len = 7))

  res.knn = classif.knn(g_mama[itrain],clas_mamaF[itrain], knn = seq(1, 9, by = 2))

  # Clasificadores regresion binaria
  res.glm = classif.glm(g_mama ~ f+l, data = lista1, basis.x = b.x1)
  res.gsam = classif.gsam(g_mama ~ s(f)+s(l), data = lista1, basis.x = b.x1)
  res.gkam = classif.gkam(g_mama ~ f+l, data = lista1, par.metric = list(X = list(metric =
      metric.lp,lp = 2)))


  # DD junto con gam
  DD.gam = classif.DD(g_mama[itrain],list(f=clas_mamaF[itrain],l=clas_mamaL[itrain]),depth="
      mode",classif="gam")
  mpr.dd = predict(DD.gam, list(test$f,test$l), type="class")

  # Predicciones del resto de clasificadores
  mpr.svmlin = predict(res.svmlin,test)
  mpr.rpart = predict(res.rpart,test)
  mpr.qda = predict(res.qda,test)
  mpr.glm = predict(res.glm,test,type="class")
```

```
57    mpr.gsam = predict(res.gsam,test,type="class")
58    mpr.gkam = predict(res.gkam,test,type="class")
59    mpr.np = predict(res.np,clas_mamaF[-itrain],type="class")
60    mpr.knn = predict(res.knn,clas_mamaF[-itrain],type="class")
61
62    g_test = g_mama[-itrain]
63
64    aux_var = calcular_tpfptnfn_mama(mpr.svmlin, g_test)
65    mr.svmlin[i,] <- unlist(aux_var)
66
67    aux_var <- calcular_tpfptnfn_mama(mpr.rpart, g_test)
68    mr.rpart[i,] <- unlist(aux_var)
69
70    aux_var <- calcular_tpfptnfn_mama(mpr.qda, g_test)
71    mr.qda[i,] <- unlist(aux_var)
72
73    aux_var <- calcular_tpfptnfn_mama(mpr.np, g_test)
74    mr.np[i,] <- unlist(aux_var)
75
76    aux_var <- calcular_tpfptnfn_mama(mpr.knn, g_test)
77    mr.knn[i,] <- unlist(aux_var)
78
79    aux_var <- calcular_tpfptnfn_mama(mpr.glm, g_test)
80    mr.glm[i,] <- unlist(aux_var)
81
82    aux_var <- calcular_tpfptnfn_mama(mpr.gsam, g_test)
83    mr.gsam[i,] <- unlist(aux_var)
84
85    aux_var <- calcular_tpfptnfn_mama(mpr.gkam, g_test)
86    mr.gkam[i,] <- unlist(aux_var)
87
88    aux_var <- calcular_tpfptnfn_mama(mpr.dd, g_test)
89    mr.dd[i,] <- unlist(aux_var)
90
91    print(paste("Iteración", i))
92  }
93
94  # Métricas de evaluación
95  m.ma.svmlin = calcular_metricas(mr.svmlin[,1], mr.svmlin[,2], mr.svmlin[,3], mr.svmlin[,4])
96  m.ma.rpart = calcular_metricas(mr.rpart[,1], mr.rpart[,2], mr.rpart[,3], mr.rpart[,4])
97  m.ma.qda = calcular_metricas(mr.qda[,1], mr.qda[,2], mr.qda[,3], mr.qda[,4])
98  m.ma.np = calcular_metricas(mr.np[,1], mr.np[,2], mr.np[,3], mr.np[,4])
99  m.ma.knn = calcular_metricas(mr.knn[,1], mr.knn[,2], mr.knn[,3], mr.knn[,4])
100 m.ma.glm = calcular_metricas(mr.glm[,1], mr.glm[,2], mr.glm[,3], mr.glm[,4])
101 m.ma.gsam = calcular_metricas(mr.gsam[,1], mr.gsam[,2], mr.gsam[,3], mr.gsam[,4])
102 m.ma.gkam = calcular_metricas(mr.gkam[,1], mr.gkam[,2], mr.gkam[,3], mr.gkam[,4])
103 m.ma.ddgam = calcular_metricas(mr.dd[,1], mr.dd[,2], mr.dd[,3], mr.dd[,4])
104
105 stat.ma.svmlin = lapply(m.ma.svmlin, calcular_stats)
106 stat.ma.rpart = lapply(m.ma.rpart, calcular_stats)
107 stat.ma.qda = lapply(m.ma.qda, calcular_stats)
108 stat.ma.np = lapply(m.ma.np, calcular_stats)
109 stat.ma.knn = lapply(m.ma.knn, calcular_stats)
110 stat.ma.glm = lapply(m.ma.glm, calcular_stats)
111 stat.ma.gsam = lapply(m.ma.gsam, calcular_stats)
112 stat.ma.gkam = lapply(m.ma.gkam, calcular_stats)
113 stat.ma.dd = lapply(m.ma.ddgam, calcular_stats)
114
115
116
117 # Cáncer de colon vs grupo de control
118 clas_colonF <- fingerprintR[grupR=='C'|grupR == 'R']
119 clas_colonL <- lipidosR[grupR=='C'|grupR == 'R']
120 g_colon <- as.factor(c(rep('C',100),rep('R',99)))
```

```
n_colon = dim(clas_colonF)[1]

# Inicializamos valores para el bucle
N=100 # número de iteraciones
r.svmlin <- matrix(NA, nrow = N, ncol = 4)
r.rpart <- matrix(NA, nrow = N, ncol = 4)
r.qda <- matrix(NA, nrow = N, ncol = 4)
r.np <- matrix(NA, nrow = N, ncol = 4)
r.knn <- matrix(NA, nrow = N, ncol = 4)
r.glm <- matrix(NA, nrow = N, ncol = 4)
r.gsam <- matrix(NA, nrow = N, ncol = 4)
r.gkam <- matrix(NA, nrow = N, ncol = 4)
r.dd <- matrix(NA, nrow = N, ncol = 4)

for (i in 1:N){
  itrain = sample(n_colon,round(n_colon*0.75),replace=FALSE)
  lista1 = list(df=data.frame(g_colon=g_colon[itrain]),f=clas_colonF[itrain],l=clas_colonL[
      itrain])
  b.x1 = list(f=create.pc.basis(clas_colonF[itrain],1:3),l=create.pc.basis(clas_colonL[
      itrain],1:2))
  test = list(df=data.frame(g_colon=g_colon[-itrain]),f=clas_colonF[-itrain],l=clas_colonL[-
      itrain])
  b.xt = list(f=create.pc.basis(clas_colonF[-itrain],1:3),l=create.pc.basis(clas_colonL[-
      itrain],1:2))

  # Clasificadores multivariantes
  res.svmlin = classif.svm(g_colon~f+l,data=lista1, kernel='linear')
  res.rpart = classif.rpart(g_colon~f+l,data=lista1)
  res.qda = classif.qda(g_colon~f+l,data=lista1)

  # Clasificación tipo kernel
  res.np = classif.np(g_colon[itrain],clas_colonF[itrain], h = seq(0.1, 0.7, len = 7))

  res.knn = classif.knn(g_colon[itrain],clas_colonF[itrain], knn = seq(1, 9, by = 2))

  # Clasificadores regresion binaria
  res.glm = classif.glm(g_colon ~ f+l, data = lista1, basis.x = b.x1)
  res.gsam = classif.gsam(g_colon ~ s(f)+s(l), data = lista1, basis.x = b.x1)
  res.gkam = classif.gkam(g_colon ~ f+l, data = lista1, par.metric = list(X = list(metric =
      metric.lp,lp = 2)))


  # DD junto con gam
  DD.gam = classif.DD(g_colon[itrain],list(f=clas_colonF[itrain],l=clas_colonL[itrain]),
      depth="mode",classif="gam")
  pr.dd = predict(DD.gam, list(test$f,test$l), type="class")

  # Predicciones del resto de clasificadores
  pr.svmlin = predict(res.svmlin,test)
  pr.rpart = predict(res.rpart,test)
  pr.qda = predict(res.qda,test)
  pr.glm = predict(res.glm,test,type="prob")
  pr.gsam = predict(res.gsam,test,type="prob")
  pr.gkam = predict(res.gkam,test,type="prob")
  pr.np = predict(res.np,clas_colonF[-itrain],type="probs")
  pr.knn = predict(res.knn,clas_colonF[-itrain],type="probs")

  g_test = g_colon[-itrain]

  aux_var = calcular_tpfptnfn_col(pr.svmlin, g_test)
  r.svmlin[i,] <- unlist(aux_var)

  aux_var <- calcular_tpfptnfn_col(pr.rpart, g_test)
  r.rpart[i,] <- unlist(aux_var)
```

```
179
180   aux_var <- calcular_tpfptnfn_col(pr.qda, g_test)
181   r.qda[i,] <- unlist(aux_var)
182
183   aux_var <- calcular_tpfptnfn_col(pr.np$group.pred, g_test)
184   r.np[i,] <- unlist(aux_var)
185
186   aux_var <- calcular_tpfptnfn_col(pr.knn$group.pred, g_test)
187   r.knn[i,] <- unlist(aux_var)
188
189   aux_var <- calcular_tpfptnfn_col(pr.glm$group.pred, g_test)
190   r.glm[i,] <- unlist(aux_var)
191
192   aux_var <- calcular_tpfptnfn_col(pr.gsam$group.pred, g_test)
193   r.gsam[i,] <- unlist(aux_var)
194
195   aux_var <- calcular_tpfptnfn_col(pr.gkam$group.pred, g_test)
196   r.gkam[i,] <- unlist(aux_var)
197
198   aux_var <- calcular_tpfptnfn_col(pr.dd, g_test)
199   r.dd[i,] <- unlist(aux_var)
200
201   print(paste("Iteración", i))
202 }
203
204 # Métricas de evaluación m.col.svmlin = calcular_metricas(mr.svmlin[,1], mr.svmlin[,2], mr.
      svmlin[,3], mr.svmlin[,4])
205 m.col.svmlin = calcular_metricas(r.svmlin[,1], r.svmlin[,2], r.svmlin[,3], r.svmlin[,4])
206 m.col.rpart = calcular_metricas(r.rpart[,1], r.rpart[,2], r.rpart[,3], r.rpart[,4])
207 m.col.qda = calcular_metricas(r.qda[,1], r.qda[,2], r.qda[,3], r.qda[,4])
208 m.col.np = calcular_metricas(r.np[,1], r.np[,2], r.np[,3], r.np[,4])
209 m.col.knn = calcular_metricas(r.knn[,1], r.knn[,2], r.knn[,3], r.knn[,4])
210 m.col.glm = calcular_metricas(r.glm[,1], r.glm[,2], r.glm[,3], r.glm[,4])
211 m.col.gsam = calcular_metricas(r.gsam[,1], r.gsam[,2], r.gsam[,3], r.gsam[,4])
212 m.col.gkam = calcular_metricas(r.gkam[,1], r.gkam[,2], r.gkam[,3], r.gkam[,4])
213 m.col.ddgam = calcular_metricas(r.dd[,1], r.dd[,2], r.dd[,3], r.dd[,4])
214
215 stat.col.svmlin = lapply(m.col.svmlin, calcular_stats)
216 stat.col.rpart = lapply(m.col.rpart, calcular_stats)
217 stat.col.qda = lapply(m.col.qda, calcular_stats)
218 stat.col.np = lapply(m.col.np, calcular_stats)
219 stat.col.knn = lapply(m.col.knn, calcular_stats)
220 stat.col.glm = lapply(m.col.glm, calcular_stats)
221 stat.col.gsam = lapply(m.col.gsam, calcular_stats)
222 stat.col.gkam = lapply(m.col.gkam, calcular_stats)
223 stat.col.dd = lapply(m.col.ddgam, calcular_stats)
```

B.3. Funciones auxiliares

```r
# FUNCIONES AUXILIARES
calcular_tpfptnfn_mama <- function(pr, test) {
  tp <- sum(test[which(pr =='M')] == 'M')
  fp <- sum(test[which(pr =='M')] == 'R')
  tn <- sum(test[which(pr =='R')] == 'R')
  fn <- sum(test[which(pr =='R')] == 'M')
  return(list(tp = tp, fp = fp, tn = tn, fn = fn))
}


calcular_tpfptnfn_col <- function(pr, test) {
  tp <- sum(test[which(pr =='C')] == 'C')
  fp <- sum(test[which(pr =='C')] == 'R')
  tn <- sum(test[which(pr =='R')] == 'R')
  fn <- sum(test[which(pr =='R')] == 'C')
  return(list(tp = tp, fp = fp, tn = tn, fn = fn))
}


calcular_f1 <- function(precision, sensibilidad) {
  f1 <- 2 * (precision * sensibilidad) / (precision + sensibilidad)
  return(f1)
}


calcular_metricas <- function(tp, fp, tn, fn){
  acc <- (tp+tn)/(tp+tn+fp+fn)
  sens <- tp/(tp+fn)
  esp <- tn/(tn + fp)
  prec <- tp/(tp+fp)
  f1 <- calcular_f1(prec,sens)
  return (list(acc=acc, sens=sens, esp=esp, prec=prec, f1=f1))
}


calcular_stats <- function(vector){
  stats <- c(mean(vector),min(vector),max(vector),sd(vector))
  return(stats)
}
```

Bibliografía

[Baker et al., 2014] Baker, M. J., Trevisan, J., Bassan, P., Bhargava, R., Butler, H. J., Dorling, K. M., Fielden, P. R., Fogarty, S. W., Fullwood, N. J., Heys, K. A., and et al. (2014). Using fourier transform ir spectroscopy to analyze biological materials. *Nature Protocols*, 9(8):1771–1791.

[Blat et al., 2019] Blat, A., Wiercigroch, E., Smeda, M., Wislocka, A., Chlopicki, S., and Malek, K. (2019). Fourier transform infrared spectroscopic signature of blood plasma in the progression of breast cancer with simultaneous metastasis to lungs. *Journal of Biophotonics*, 12(10).

[Cuesta-Albertos and Febrero-Bande, 2010] Cuesta-Albertos, J. A. and Febrero-Bande, M. (2010). A simple multiway anova for functional data. *TEST*, 19(3):537–557.

[Cuevas et al., 2006] Cuevas, A., Febrero, M., and Fraiman, R. (2006). On the use of the bootstrap for estimating functions with functional data. *Computational Statistics Data Analysis*, 51(2):1063–1074.

[Cuevas et al., 2007] Cuevas, J. A., Febrero, M., and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22(3):481–496.

[Febrero et al., 2007] Febrero, M., Galeano, P., and González-Manteiga, W. (2007). Outlier detection in functional data by depth measures, with application to identify abnormal nox levels. *Environmetrics*, 19(4):331–345.

[Febrero-Bande and Oviedo de la Fuente, 2012] Febrero-Bande, M. and Oviedo de la Fuente, M. (2012). Statistical Computing in Functional Data Analysis: The R Package fda.usc. *Journal of Statistical Software*, 51(4):1–28.

[Ferraty and Vieu, 2006] Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer New York: Springer e-books.

[Fraiman and Muniz, 2001] Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *Test*, 10(2):419–440.

[García Díaz, 2023] García Díaz, S. (2023). Diseño y simulación de antenas metálicas para aplicaciones biológicas. Trabajo fin de máster, Universidad de Oviedo, Oviedo. Master en Física Avanzada, Especialidad en Nanofísica y Materiales Cuánticos.

[Guang et al., 2020] Guang, P., Huang, W., Guo, L., Yang, X., Huang, F., Yang, M., Wen, W., and Li, L. (2020). Blood-based ftir-atr spectroscopy coupled with extreme gradient boosting for the diagnosis of type 2 diabetes. *Medicine*, 99(15).

[Huber et al., 2021] Huber, M., Kepesidis, K. V., Voronina, L., Fleischmann, F., Fill, E., Hermann, J., Koch, I., Milger-Kneidinger, K., Kolben, T., Schulz, G. B., Jokisch, F., Behr, J., Harbeck, N., Reiser, M., Stief, C., Krausz, F., and Zigman, M. (2021). Infrared molecular fingerprinting of blood-based liquid biopsies for the detection of cancer. *eLife*, 10:e68758.

[Jayasinghe et al., 2023] Jayasinghe, M., Prathiraja, O., Caldera, D., Jena, R., Coffie-Pierre, J. A., Silva, M. S., and Siddiqui, O. S. (2023). Colon cancer screening methods: 2023 update. *Cureus*.

[Kepesidis et al., 2021] Kepesidis, K. V., Bozic-Iven, M., Huber, M., Abdel-Aziz, N., Kullab, S., Abdelwarith, A., Al Diab, A., Al Ghamdi, M., Hilal, M. A., Bahadoor, M. R., and et al. (2021). Breast-cancer detection using blood-based infrared molecular fingerprints. *BMC Cancer*, 21(1).

[Navarro et al., 2017] Navarro, M., Nicolas, A., Ferrandez, A., and Lanas, A. (2017). Colorectal cancer population screening programs worldwide in 2016: An update. *World Journal of Gastroenterology*, 23(20):3632.

[Paraskevaidi et al., 2017] Paraskevaidi, M., Morais, C. L., Lima, K. M., Snowden, J. S., Saxon, J. A., Richardson, A. M., Jones, M., Mann, D. M., Allsop, D., Martin-Hirsch, P. L., and et al. (2017). Differential diagnosis of alzheimer's disease using spectrochemical analysis of blood. *Proceedings of the National Academy of Sciences*, 114(38).

[Ramsay and Silverman, 2005] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Science+Business Media, Inc.

[Sung et al., 2021] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249.

[Warner, 2011] Warner, E. (2011). Breast-cancer screening. *New England Journal of Medicine*, 365(11):1025–1032.

[Wong et al., 2015] Wong, M. C., Ching, J. Y., Chan, V. C., Lam, T. Y., Luk, A. K., Ng, S. S., and Sung, J. J. (2015). Factors associated with false-positive and false-negative fecal immunochemical test results for colorectal cancer screening. *Gastrointestinal Endoscopy*, 81(3):596–607.